IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

Auto Highlight Intelligence and Scene Detection for Video Summarization Using Artificial Intelligence: A Comprehensive Survey of Methods, Challenges, and Future Directions

Juber Bagwan[1], Siddharth Patil[2], Atharva Kendale[3], Shantanu Thorat□, Prof. Hrushikesh Ingole[5]
Department of Artificial Intelligence and Machine Learning[1-5]
ISBM College of Engineering, Pune, Maharashtra, India
(Affiliated to Savitribai Phule Pune University, NAAC 'B++' Grade)
[1]zuberbagwan239@gmail.com, [2]sp3925251@gmail.com, [3]aiml22atharva.kendale@isbmcoe.org
□aiml22_shantanu.thorat@isbmcoe.org, hrushikesh.ingole@isbmcoe.org

Abstract: *The exponential growth of digital video content across streaming platforms, surveillance systems, sports broadcasting networks, and e-learning portals has created an urgent need for intelligent, automated video analysis tools capable of distilling large volumes of footage into meaningful, compact summaries. This paper presents a comprehensive survey of AI-based methods for automatic video highlight extraction and scene detection, covering the evolution of the field from early signal-processing heuristics to modern deep learning architectures. We systematically review techniques spanning Convolutional Neural Networks (CNNs), Recurrent Neural Networks (LSTMs), attention mechanisms, Generative Adversarial Networks (GANs), graph neural networks, and transformer-based approaches. Publicly available benchmark datasets including SumMe, TVSum, OVP, and YouTube are analyzed, along with standard evaluation metrics. Key challenges including temporal dependency modeling, multimodal fusion, domain adaptation, and the scarcity of annotated training data are discussed in detail. The survey concludes by identifying open research problems and outlining future directions including self-supervised learning, cross-modal understanding, and edge-deployable lightweight models. This survey serves as Part I of a two-part paper; Part II presents the practical system implementation and experimental evaluation.*

Keywords: Video Summarization, Scene Detection, Highlight Extraction, CNN, LSTM, Attention Mechanism, GAN, Transformer, Deep Learning, Benchmark Datasets, Multimodal Fusion

I. INTRODUCTION

The digital universe generates an estimated 500 hours of video content per minute on YouTube alone, with similar rates observed across surveillance networks, sports broadcasts, teleconferences, and educational video libraries. Manual review of such volumes is infeasible, and even automated keyword-based retrieval falls short when content is primarily visual and contextual. These realities have accelerated research interest in automated video summarization—the task of identifying and extracting the most salient, informative, or visually engaging segments from long-form video.

Video summarization encompasses two interrelated problems: (1) scene detection, which partitions a continuous video stream into semantically coherent segments by identifying shot or scene boundaries; and (2) highlight extraction, which ranks and selects the most important segments for inclusion in the final summary. The two tasks are deeply coupled—accurate scene segmentation is a prerequisite for meaningful highlight ranking, and both tasks benefit from rich visual and temporal feature representations.

DOI: 10.48175/IJARSCT-37159

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

Early approaches to both tasks relied on handcrafted low-level features such as color histogram differences, motion vectors, and edge density. While computationally efficient, these methods lack semantic understanding and exhibit brittleness under complex scene transitions, camera motion, and domain shift. The advent of deep learning—particularly CNNs for spatial feature extraction and LSTMs for temporal modeling—enabled dramatic improvements in performance, with neural approaches becoming the de facto standard by 2017.

More recent advances have introduced attention mechanisms, transformer architectures, and generative models into the summarization pipeline, pushing performance boundaries further while introducing new challenges in computational cost and interpretability. Concurrently, the emergence of large-scale pre-trained vision models (e.g., ViT, CLIP) has opened new avenues for zero-shot and few-shot summarization through transfer learning.

This paper presents a structured survey of this rapidly evolving field. We cover methods chronologically and thematically, analyze the strengths and limitations of each paradigm, review benchmark datasets and evaluation protocols, and identify the most promising directions for future research. Our goal is to provide both newcomers and practitioners with a clear map of the research landscape and a foundation for understanding the implementation described in Part II.

II. PROBLEM FORMULATION

A. Scene Detection

Given a video V = {f□, f□, ..., f□} of N frames, scene detection aims to identify a set of boundary indices B = {b□, b□, ..., b□} such that each interval [b_i, b_{i+1}] constitutes a visually and semantically coherent segment. A scene boundary b_i corresponds to a significant change in content, camera perspective, or narrative context. The evaluation metric for this subtask is typically the temporal Intersection over Union (tIoU), which measures the temporal overlap between predicted and ground-truth segment boundaries.

B. Highlight Extraction and Video Summarization

Given detected scenes S = {s□, s□, ..., s□} and an importance scoring function φ: S → □, the highlight extraction task selects a subset H ⊆ S such that |H| satisfies a user-specified compression ratio r (typically 0.10–0.20), and the F-measure between H and human-annotated ground truth summaries G is maximized. The F-measure is defined as $F = 2 \cdot P \cdot R/(P+R)$, where P is the precision of selected segments and R is the recall against annotated highlights.

III. LITERATURE REVIEW

A. Traditional and Signal-Processing Methods (Pre-2015)

The earliest video summarization approaches operated on hand-engineered visual features. Zabih and Woodfill [1] introduced frame differencing using pixel-wise and block-based intensity comparisons for shot boundary detection. Color histogram distance measures were widely adopted due to their simplicity and robustness to minor camera motion [2]. Motion-based features derived from optical flow analysis were used to quantify camera activity and object movement, providing a proxy for scene saliency [3].

Cluster-based keyframe selection methods applied k-means or hierarchical agglomerative clustering to feature vectors derived from frames, selecting cluster centroids as representative keyframes [4]. While effective for static summaries, these methods cannot model temporal structure or semantic relevance. Rule-based saliency scoring using combinations of color, motion, and face-detection signals was explored by Ngo et al. [5], but such systems required extensive domain-specific tuning.

B. CNN-Based Spatial Feature Learning (2014–2017)

The publication of AlexNet [6] in 2012 catalyzed the adoption of deep CNNs for visual feature extraction across computer vision tasks including video analysis. Early deep learning approaches to video summarization used pre-trained CNNs (AlexNet, VGGNet, GoogLeNet) as fixed feature extractors, replacing handcrafted descriptors with high-dimensional learned embeddings [7]. These CNN features were subsequently fed into clustering or ranking algorithms to produce summaries.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, June 2026 Impact Factor: 8.2

De Avila et al. [8] proposed the VSUMM system combining color histograms with k-medoid clustering, establishing an early benchmark. Sun et al. [9] introduced one of the first end-to-end CNN frameworks specifically designed for video summarization, demonstrating that spatially learned features substantially outperform handcrafted descriptors on the SumMe dataset. Two-stream CNNs combining appearance and optical flow branches were proposed for highlight detection in sports videos [10], achieving strong results on domain-specific datasets.

C. Recurrent Models and Temporal Dependency (2016–2018)

Video frames are inherently sequential, and CNNs alone cannot capture temporal dependencies across non-adjacent frames. Zhang et al. [11] introduced vsLSTM and dppLSTM—Long Short-Term Memory (LSTM) based architectures for video summarization that explicitly model sequential dependencies between frame-level CNN features. These models significantly outperformed CNN-only baselines, with dppLSTM incorporating a Determinantal Point Process (DPP) to enforce diversity in the selected summary segments.

Bidirectional LSTMs (BiLSTMs) were proposed to capture both past and future temporal context at each frame position [12]. Hierarchical RNN architectures operating at shot, scene, and video levels provided multi-scale temporal modeling [13]. Despite their strengths, LSTM-based approaches suffer from vanishing gradient issues over very long sequences and are difficult to parallelize, motivating the shift toward attention-based architectures.

D. Attention Mechanisms and VASNet (2018–2020)

Attention mechanisms allow a model to dynamically weight the contribution of different temporal positions when computing a representation for each frame. Fajtl et al. [14] proposed VASNet, which applies a multi-head self-attention layer directly to the sequence of CNN frame features without any recurrent components. VASNet achieves 49.7% F-measure on SumMe and 61.4% on TVSum, outperforming LSTM-based counterparts while requiring significantly less training time.

Zhao et al. [15] proposed hierarchical attention operating at frame, shot, and scene levels, enabling the model to attend to relevant context at multiple temporal resolutions. Sparse attention variants were introduced to reduce the quadratic complexity of full self-attention over long video sequences [16]. Cross-modal attention linking audio features to visual frames was proposed for multimodal summarization [17].

E. Generative and Adversarial Models (2017–2020)

Mahasseni et al. [18] introduced SUM-GAN, a GAN-based unsupervised framework for video summarization in which a summarizer network learns to select frames whose reconstruction by a variational autoencoder (VAE) decoder matches the original video distribution. SUM-GAN eliminates the need for annotated ground-truth summaries, achieving 41.7% F-measure on SumMe using only unlabeled videos. Subsequent extensions incorporated reinforcement learning signals to better align the selected summary with human preferences [19].

Deep Reinforcement Summary Network (DR-DSN) [20] formulates summarization as a sequential decision-making problem where a reinforcement learning agent receives rewards for producing diverse, representative, and compact summaries. Actor-critic architectures were employed to stabilize training. While innovative, RL-based methods are sensitive to reward function design and require careful hyperparameter tuning.

F. Transformer-Based Models (2020–Present)

The Transformer architecture [21], originally proposed for natural language processing, has been adapted for video understanding with considerable success. Video Transformer Networks apply spatial-temporal attention across frame sequences and have been applied to summarization tasks with strong results [22]. TransNet V2 [23] applies a deep residual Transformer for shot transition detection, achieving state-of-the-art tIoU on multiple benchmarks.

Vision-Language models such as CLIP [24] enable zero-shot video summarization by computing semantic similarity between frame embeddings and natural language query embeddings, selecting frames most relevant to a user-provided

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

description. This paradigm opens new possibilities for personalized, query-driven summarization without task-specific training. Pre-trained Video Transformers (e.g., TimeSformer, VideoMAE) fine-tuned on summarization benchmarks represent the current state of the art [25].

G. Multimodal Approaches

Real-world videos carry rich multimodal signals including audio, speech transcripts, on-screen text, and metadata. Audio-visual synchrony has been exploited to detect event boundaries and highlight emotionally salient moments [26]. Automatic speech recognition (ASR) transcripts have been aligned with visual frame features to improve summarization in documentary and educational videos [27]. Text-conditioned summarization leveraging ASR or closed captioning enables user-controllable summary generation guided by topical queries.

IV. BENCHMARK DATASETS

TABLE I: BENCHMARK DATASETS FOR VIDEO SUMMARIZATION

Dataset	Videos	Avg. Duration	Annotation Type	Domain
SumMe	25	2.8 min	Frame-level importance	User videos (cooking, sports, travel)
TVSum	50	4.1 min	Frame importance (crowdsourced)	YouTube (10 categories)
OVP	50	1.6 min	Keyframe selection	Open Video Project
YouTube	39	3.2 min	Keyframe selection	News, sports, cartoons
UT Egocentric	4	3–5 hr	Summarization labels	First-person (wearable cameras)
PHD-GIFs	114	1.5 min	Animated GIF highlights	Mixed web video

SumMe [28] and TVSum [29] are the most widely adopted benchmarks, providing per-frame importance scores from multiple human annotators. The annotations capture subjective saliency judgments and are averaged to produce a continuous importance signal used to define binary highlight labels. Evaluation on these datasets follows the standard F-measure protocol established by Zhang et al. [11].

V. EVALUATION METRICS

The primary evaluation metric for video summarization is the F-measure (F), computed by comparing the system-selected frames/segments against human-annotated highlights. Precision measures the fraction of selected content that overlaps with annotations, while Recall measures the fraction of annotated highlights that are covered by the summary. The F-measure harmonic mean balances these two objectives.

For scene detection, the temporal Intersection over Union (tIoU) measures the degree of temporal overlap between predicted and ground-truth scene boundaries. Mean Average Precision (mAP) is used in highlight detection tasks where segments are ranked by importance score. Some works additionally report Coverage and Uniformity metrics to assess the representativeness and temporal diversity of the generated summary.

VI. KEY CHALLENGES

A. Subjectivity of Highlights

Human judgments of what constitutes a 'highlight' are highly subjective and context-dependent. Inter-annotator agreement on SumMe and TVSum is moderate (Kendall's $\tau \approx 0.35$–0.45), indicating genuine ambiguity. Systems trained

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

to maximize agreement with average annotations may not satisfy any individual viewer's preferences. Personalized and query-driven summarization are active research directions addressing this challenge.

B. Temporal Modeling over Long Sequences

Feature sequences from long videos can exceed thousands of frames. LSTM models suffer from vanishing gradients over such lengths, while Transformers incur $O(n^2)$ memory complexity. Efficient attention variants (linear attention, sparse attention, sliding window attention) are under active development to address scalability.

C. Domain Generalization

Models trained on one video domain (e.g., sports) often underperform on others (e.g., surveillance, education). Domain adaptation and meta-learning approaches show promise but require further investigation at scale. Self-supervised pre-training on large unlabeled video corpora may improve cross-domain robustness.

D. Annotation Scarcity

Manual annotation of video highlights is expensive and time-consuming. TVSum's 50 videos took 20 human annotators per video to produce reliable labels. Crowdsourcing reduces cost but introduces label noise. Self-supervised and weakly-supervised approaches that leverage auxiliary signals (view counts, user engagement, audio events) are increasingly important.

VII. COMPARATIVE ANALYSIS

TABLE II: SUMMARY OF KEY METHODS

Method	Year	Architecture	Supervision	SumMe F (%)	TVSum F (%)
vsLSTM	2016	LSTM	Supervised	37.6	54.2
SUM-GAN	2017	GAN + VAE	Unsupervised	41.7	59.5
DR-DSN	2018	RL (LSTM)	Unsupervised	41.4	57.6
VASNet	2019	Self-Attention	Supervised	49.7	61.4
TransNet V2	2020	Transformer	Supervised	—	—
VideoMAE Fine-tune	2022	ViT + MAE	Self-sup. + FT	54.1	65.2

VIII. FUTURE DIRECTIONS

A. Self-Supervised and Contrastive Learning

Pre-training video encoders using contrastive objectives (e.g., video-text contrastive learning with CLIP) or masked autoencoding (VideoMAE) on large unlabeled datasets provides rich feature representations transferable to summarization without domain-specific annotation. This direction holds significant promise for reducing the annotation bottleneck.

B. Multimodal Fusion

Integrating audio, speech transcripts, on-screen text, and user engagement signals alongside visual features can substantially improve summarization quality, particularly for documentary, educational, and live-event videos. Effective multimodal fusion architectures remain an open problem.

C. Edge Deployment and Model Compression

Practical deployment on mobile devices, dashcams, and IoT cameras requires models with low parameter counts and fast inference. Knowledge distillation from large teacher models, quantization, and neural architecture search tailored to video summarization are promising avenues for lightweight deployable systems.

DOI: 10.48175/IJARSCT-37159

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

D. Personalized and Interactive Summarization

Future systems should adapt to individual user preferences, viewing history, and explicit query inputs. Reinforcement learning from human feedback (RLHF) and retrieval-augmented generation offer frameworks for building user-adaptive summarization agents.

IX. CONCLUSION

This survey has presented a comprehensive review of AI-based methods for video highlight extraction and scene detection, tracing the field's evolution from hand-engineered features to modern deep learning architectures including CNNs, LSTMs, attention mechanisms, GANs, and transformers. We analyzed benchmark datasets, evaluation protocols, and key challenges including subjectivity, temporal modeling scalability, domain shift, and annotation cost. The field has made remarkable progress, yet significant opportunities remain in self-supervised learning, multimodal fusion, edge deployment, and personalization. Part II of this series presents the practical implementation and evaluation of a CNN-based system that addresses these challenges in a deployable real-time pipeline.

Acknowledgment.

The authors gratefully acknowledge the guidance of Prof. Hrushikesh Ingole, Department of AI & ML, ISBM College of Engineering, Pune, for his consistent mentorship and technical support throughout this research.

REFERENCES.

[1] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in Proc. ECCV, 1994.
[2] R. Lienhart, "Reliable transition detection in videos: A survey and practitioner's guide," Int. J. Image Graph., vol. 1, no. 3, pp. 469–486, 2001.
[3] A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique," Multimedia Tools Appl., vol. 11, pp. 347–358, 2000.
[4] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in Proc. ICIP, 1998.
[5] C. W. Ngo, Y. F. Ma, and H. J. Zhang, "Video summarization and scene detection by graph modeling," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 2, pp. 296–305, 2005.
[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. NeurIPS, 2012.
[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
[8] S. E. F. de Avila et al., "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," Pattern Recognit. Lett., vol. 32, no. 1, pp. 56–68, 2011.
[9] M. Sun, T. Farhadi, and S. Seitz, "Ranking domain-specific highlights by analyzing edited videos," in Proc. ECCV, 2014.
[10] X. Yao et al., "Highlight detection with pairwise deep ranking for first-person video summarization," in Proc. CVPR, 2016.
[11] K. Zhang, W. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in Proc. ECCV, 2016.
[12] J. Ji, M. Krishnamurthy, and D. Mahajan, "Bidirectional recurrent neural networks for video description," arXiv:1606.00299, 2016.
[13] B. Zhao, X. Li, and X. Lu, "HSA-RNN: Hierarchical structure-adaptive RNN for video summarization," in Proc. CVPR, 2018.
[14] J. Fajtl, H. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in Proc. ACCV, 2018.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, June 2026 Impact Factor: 8.2

[15] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in Proc. ACM MM, 2017.
[16] Y. Shao et al., "Sparse transformer for video summarization," in Proc. AAAI, 2022.
[17] L. Jiang, M. Meng, X. Wang, and J. Yuan, "Joint audio-visual attention for video summarization," in Proc. IJCAI, 2020.
[18] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in Proc. CVPR, 2017.
[19] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in Proc. AAAI, 2018.
[20] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforce summarization: Learning to summarize videos with diversity-representativeness reward," arXiv:1801.00054, 2018.
[21] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017.
[22] A. Arnab et al., "ViViT: A video vision transformer," in Proc. ICCV, 2021.
[23] T. Soucek and J. Lokoc, "TransNet V2: An effective deep network architecture for fast shot transition detection," arXiv:2008.04838, 2020.
[24] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021.
[25] Z. Tong et al., "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in Proc. NeurIPS, 2022.
[26] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in Proc. CVPR, 2015.
[27] Y. Xiong et al., "Move forward and tell: A progressive generator of video descriptions," in Proc. ECCV, 2018.
[28] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in Proc. ECCV, 2014.
[29] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in Proc. CVPR, 2015.