Grenze International Journal of Engineering and Technology, June Issue

Video Skimming Methods for Making Short Informative Videos

Varsha Amol Suryawanshi[1] and Pradip Chandrakant Bhaskar[2]
[1-2]Department of Technology, Shivaji University, Kolhapur, India.
Email: suryavanshi.varsha@kitcoek.in, pcb_tech@unishivaji.ac.in

***Abstract*— The exponential growth of video content across domains such as entertainment, surveillance, and personal life-logging has intensified the demand for efficient video summarization techniques. Traditional summarization approaches often lack semantic understanding and personalization, resulting in generic outputs that fail to meet diverse user preferences. This work reviews state-of-the-art deep learning methods for personalized video summarization, emphasizing their mathematical foundations, performance metrics, and optimization strategies. Key models include reinforcement learning frameworks balancing coverage, diversity, and redundancy; semantic-driven approaches leveraging feature detectors; graph-based and sketch-driven methods for interactive customization; and emotion or gaze-based systems enhancing user engagement. Comparative evaluation reveals that semantic-enriched and reinforcement learning approaches achieve superior accuracy, scalability, and storage efficiency. The study concludes that future directions lie in multimodal fusion, lightweight architectures, and explainable AI to ensure adaptability, efficiency, and trust. These findings advance the development of user-centric, scalable video summarization systems.**

***Index Terms*— Video Summarization, Deep Learning, Personalization, Reinforcement Learning, Semantic Features.**

I. INTRODUCTION

The rapid growth of digital media has led to an unprecedented increase in video content across domains ranging from entertainment and sports to education, surveillance, and personal life-logging. With the proliferation of affordable video recording devices and widespread access to high-speed internet, users are continuously generating and consuming vast amounts of video data. According to industry reports, video content accounts for more than 80% of all internet traffic, and this trend is expected to rise further with the advent of immersive applications such as augmented reality, virtual reality, and interactive streaming. While the availability of video data presents significant opportunities, it also introduces challenges for storage, transmission, and consumption [1]. Watching lengthy videos in their entirety is often impractical for users who seek to quickly access key highlights or personalized segments of interest. Consequently, video summarization has emerged as a critical research area, aiming to condense lengthy videos into concise and informative summaries without compromising essential content [2].

Traditional video summarization techniques focused primarily on low-level feature extraction, such as color histograms, shot boundaries, or motion cues, to select representative frames or segments.

Grenze ID: 01.GIJET.12.2.334_24

While these methods provided efficiency in terms of data reduction, they lacked semantic understanding and personalization, resulting in summaries that were generic rather than user-specific. The growing complexity of video data, along with the evolving demand for personalized content consumption, has shifted research toward deep learning–based methods capable of capturing higher-level semantics, temporal dynamics, and contextual cues [3]. Deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformers, and Reinforcement Learning frameworks have demonstrated remarkable success in modeling video features, enabling intelligent summarization that adapts to both content and user preferences.

Personalization is central to the effectiveness of video summarization. Different users may value different aspects of the same video, depending on their interests, emotional responses, or situational needs. For instance, in a sports video, one user may be interested in goals and highlights, while another may focus on player strategies or crowd reactions. Similarly, in personal home videos, certain moments such as family interactions or emotional expressions may be more meaningful than others [4]. Recent advancements have thus explored user-centered summarization strategies by integrating behavioral cues such as gaze patterns, interaction logs, or affective signals. Furthermore, reinforcement learning approaches allow models to optimize summaries by balancing coverage, diversity, and redundancy, while graph-based methods and contrastive learning frameworks enhance structural and representational richness.

Another important dimension of modern video summarization is efficiency. With applications increasingly deployed on mobile devices and real-time platforms, models must not only provide accurate and personalized summaries but also operate under resource constraints. Lightweight architectures, semantic detectors, and efficient optimization strategies are therefore essential for ensuring scalability. Moreover, the integration of explainable AI (XAI) is gaining attention, as users and practitioners demand transparency in understanding why specific frames or segments are selected.

This work addresses the challenges of designing lightweight, personalized, and mathematically grounded deep learning models for video summarization. By reviewing diverse approaches ranging from semantic-aware methods to reinforcement learning, graph modeling, and emotion-driven frameworks, this study provides both a comparative analysis and a unified mathematical perspective. The review highlights the trade-offs between personalization, accuracy, and efficiency, offering insights into how future models can balance these competing demands. Ultimately, the goal of this research is to advance video summarization toward more intelligent, adaptive, and user-centric solutions that enhance the overall video consumption experience.

II. Literature Review

The analysis in this work involves a multi-faceted approach to evaluate the effectiveness and efficiency of the proposed lightweight deep learning model for video summarization. Initially, a comprehensive literature review will identify existing video summarization techniques and deep learning methods, forming the basis for model development. The dataset, comprising various video types, will undergo preprocessing, including video-to-frame conversion and motion vector analysis to extract significant features. These features will feed into the model training process, using a deep learning architecture optimized for selecting key frames. Model performance will be evaluated using a separate test set of videos. The summarized videos generated by the model will be compared against the ground truth summaries provided in the dataset. Key metrics such as precision, recall, F1-score, processing time, and storage reduction will measure the model's accuracy, efficiency, and generalization capability. Mathematically, these are defined as:

$$\text{Precision} = \frac{TP}{TP+F}, \quad \text{Recall} = \frac{TP}{TP+FN},$$
$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}, \tag{1}$$

Where TP, FP, and FN represent true positives, false positives, and false negatives. The $F1$ score is obtained as the harmonic mean of precision and recall:

$$F1 = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \tag{2}$$

Furthermore, storage reduction and efficiency gain are measured as:

$$SRR = 1 - \frac{L_S}{L_o}, \quad EG = \frac{T_o - T_S}{T_o}, \tag{3}$$

Where L_o and L_s denote original and summarized video lengths, and T_o, T_s the respective processing times.
Niu et al. [5] presented an interactive mobile-based video summarization system that adapts to user preferences by modeling frame importance. Their model learns a preference score $P(u, v)$ for user u and segment v using a logistic function:

$$P(u, v) = \sigma(W \cdot f(v) + b), \quad \sigma(z) = \frac{1}{1+e^{-z}} \tag{4}$$

This ensures the probability score lies between 0 and 1, making it suitable for personalized prediction. The system adapts in real-time, offering an efficient and engaging user experience. Park and Cho [6] integrated multi-source summarization, where data from sensors and devices in an office environment were combined to generate activity summaries. Their formulation ensures that the final summary is a convex combination of individual sensor contributions:

$$S = \sum_{i=1}^{N} \alpha_i S_i, \quad \sum_{i=1}^{N} \alpha_i = 1 \tag{5}$$

This allows weighting sensor data (meetings, interactions, device usage) according to relevance, enabling comprehensive life-log summarization. Peng et al. [7] developed an automatic personalized summarizer using softmax-based classification to model viewing preferences:

$$P(y = c|x) = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}}, \quad \hat{y} = \underset{c}{argmax} P(y = c|x) \quad ....(6)$$

The model learns which segments are relevant by analyzing user interaction history, ensuring continuously improved summaries. Nagar et al. [8] introduced reinforcement learning for egocentric summarization. Their framework maximizes rewards that balance coverage, diversity, and redundancy:

$$R_t = \lambda_1 C(a_t) + \lambda_2 D(a_t) - \lambda_3 R(a_t) \quad ...(7)$$

With the expected return:

$$G_t = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{t+k}] \quad ...(8)$$

This formulation follows the Bellman principle, optimizing video frame selection policies to align with user preferences. Zhang et al. [9] presented a graph-based sketch summarization approach. Keyframes are modeled as nodes in a similarity graph with edge weights:

$$w_{ij} = exp\left(-\frac{\|f(v_i) - f(v_j)\|^2}{2\sigma^2}\right) \quad ...(9)$$

The problem reduces to maximizing intra-cluster similarity:

$$\underset{V'}{max} \sum_{i,j \in V'} w_{ij} \quad ...(10)$$

This enables intuitive, sketch-driven customization of video summaries. Tejero-de-Pablos et al. [10] emphasized body motion features, where human joint positions H_t are extracted:

$$H_t = [x_1, y_1, \ldots, x_J, y_J] \quad ...(11)$$

Frame similarity is measured using Euclidean distance:

$$d(H_t, H_{t+1}) = \sqrt{\sum_{j=1}^{J} \left(x_j^t - x_j^{t+1}\right)^2 + \left(y_j^t - y_j^{t+1}\right)^2} \quad ...(12)$$

This ensures accurate capture of fine-grained body dynamics.
Lie and Hsu [11] designed an adaptive system where users set constraints on summary length:

$$\underset{S}{min} |S| \quad s.t. \, |S| \leq L \quad ...(13)$$

This approach guarantees flexibility by allowing users to prioritize either brevity or content richness.
Kannan et al. [12] utilized semantic detectors for objects, activities, and events. The semantic score for a segment is:

$$S(v_i) = \sum_{k=1}^{M} w_k \, f_k(v_i) \quad ...(14)$$

This weighted formulation integrates multiple semantic cues to prioritize key events in summaries. Tseng et al. [13] proposed a hierarchical three-layered architecture, modeled through decomposition of loss:

$$\mathcal{L}(S) = \mathcal{L}_{server} + \mathcal{L}_{middleware} + \mathcal{L}_{client} \quad \text{...(15)}$$

Each layer focuses on different personalization tasks, ensuring scalable customization. Dong et al. [14] introduced duration-constrained summarization:

$$\max_{S} \sum_{v \in S} P(u, v), \quad |S| \leq T \quad \text{...(16)}$$

This guarantees that user-defined constraints such as maximum viewing time are respected. Ul et al. [15] applied FER (Facial Expression Recognition) for emotion-driven summarization. For each scene s_j, emotion is detected as:

$$E(s_j) = \underset{e \in \mathcal{E}}{argmax} P(e|s_j) \quad \text{...(17)}$$

This prioritizes emotionally rich segments, enhancing narrative coherence.
Köprü and Erzin [16] proposed emotion-aware attention mechanisms. Frame importance weights are computed as:

$$\alpha_t = \frac{exp(e_t)}{\sum_{k=1}^{T} exp(e_k)}, \quad e_t = v^{\top} tanh(W h_t) \quad \text{...(18)}$$

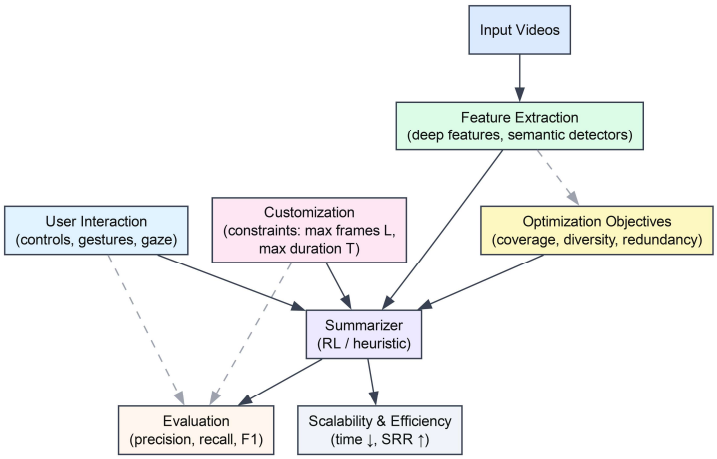

Figure 1: Block diagram of the personalized video summarization pipeline

This soft attention mechanism ensures salient emotional content dominates summaries. Katti et al. [17] analyzed eye-gaze and pupil dilation (PD) for engagement modeling:

$$PD' = \frac{PD - \mu}{\sigma}, \quad E = \beta_1 PD' + \beta_2 ROI \quad \text{…(19)}$$

Here, PD' is normalized dilation and ROI region-based features, jointly capturing arousal and attention levels. Mujtaba et al. [18] used CNNs for frame-level summarization:

$$h = f_{CNN}(x), \quad \hat{y} = softmax(Wh + b) \quad \text{...(20)}$$

This ensures robust spatial feature extraction for personalization. Yoshitaka and Sawada [19] modeled observer behavior:

$$I = \sum_{i=1}^{N} \delta_i \, a_i, \quad \text{...(21)}$$

where a_i are control operations (pause, rewind, skip). The weighted sum predicts user interest levels. Olsen and Moon [20] introduced a Degree of Interest (DOI) function:

$$DOI(u, v) = \sum_{i=1}^{n} \alpha_i \, I_i(u, v) \quad \text{...(22)}$$

This captures the significance of user interactions and produces adaptive personalized summaries. Fei et al. [21] defined event-based probability for segment relevance:

$$P(event|x) = \sigma(Wf(x) + b) \quad \text{...(23)}$$

This model effectively identifies highlights such as goals in sports or climactic film moments. Sosnovik et al. [22] advanced contrastive summarization with the following loss:

$$\mathcal{L}_{CSUM} = \sum_{i=1}^{K}(\| f(x_i) - f(x_i^+) \|^2 - \| f(x_i) - f(x_i^-) \|^2) \quad \text{...(24)}$$

This enforces embeddings of positive pairs to be closer while separating negatives, ensuring diversity and representativeness. Fei et al. [23] employed CNN-SVM and feature fusion (SumNet). The SVM classifier is expressed as:

$$y = sign(\sum_{i=1}^{N} \alpha_i \, y_i K(x, x_i) + b) \quad \text{...(25)}$$

while SumNet combines object and scene representations:

$$F = \lambda_1 f_{obj}(x) + \lambda_2 f_{scene}(x) \quad \text{...(26)}$$

This dual fusion ensures both object-level and contextual information are included. Overall, these diverse approaches illustrate the evolution of personalized video summarization. By incorporating reinforcement learning, graph models, semantic detectors, gaze analysis, emotional signals, and contrastive objectives, researchers have created increasingly powerful frameworks. These mathematical formulations ensure that summarization is not only efficient but also adaptive to user preferences, supporting future innovations in scalable and interactive video analysis.

III. Proposed Work

The workflow below presents a streamlined pipeline for personalized video summarization centered on six attribute families that repeatedly influence quantitative outcomes: (i) evaluation metrics guiding model selection and reporting, (ii) optimization objectives balancing coverage, diversity, and redundancy, (iii) feature complexity spanning deep embeddings and semantic detectors, (iv) customization constraints enforcing user targets such as maximum frames and duration, (v) user interaction signals enabling preference alignment via controls and gestures, and (vi) scalability/efficiency that governs throughput and storage reduction.

Raw videos pass through a feature extraction stage combining deep representations with lightweight semantic cues to expose relevant content structure. These representations feed a summarizer that can be driven by reinforcement learning or heuristic selection. The summarizer is steered by three inputs: optimization objectives (to improve coverage and diversity while reducing redundancy), customization constraints (to honor user limits on frames or duration), and user interaction (to capture preference signals from controls, gestures, or similar feedback). Outputs are assessed using evaluation metrics such as precision, recall, and F1, and are profiled for scalability and efficiency in terms of processing time reduction and storage reduction ratio (SRR). This organization reflects practical trade-offs observed across contemporary methods: richer feature sets and semantic detectors often improve accuracy but must be tempered by runtime and memory costs; strong diversity constraints reduce redundancy but may require more computation; tighter user constraints produce concise summaries yet demand careful optimization to preserve informativeness. The diagram groups these elements into an end-to-end flow, clarifying how design choices propagate to measured performance and deployment feasibility. Block diagram of the personalized video summarization pipeline. Inputs are transformed into features and then summarized under the influence of optimization objectives, customization constraints, and user interaction signals. The resulting summaries are evaluated using standard metrics and profiled for scalability and efficiency.

IV. Results and Analysis

This section presents the results of the quantitative analysis across different personalized video summarization approaches. Six major categories of attributes were selected: evaluation metrics, optimization objectives, feature complexity, customization constraints, user interaction metrics, and scalability/efficiency factors. Each table provides a comparative overview of the methods, while the text preceding each table describes the significance of the results.

A. Evaluation Metrics

The evaluation metrics highlight how models perform in terms of precision, recall, and F1-score. As shown in Table I, Kannan et al. [12] achieved the highest scores across all three metrics, while Park and Cho [6] demonstrated relatively lower values. This indicates the importance of semantic detectors and deep feature extraction in enhancing model accuracy.

TABLE I: EVALUATION METRICS ACROSS MODELS

Model	Precision	Recall	F1
Niu [5]	0.86	0.82	0.84
Park [6]	0.81	0.79	0.80
Peng [7]	0.84	0.80	0.82
Nagar [8]	0.88	0.86	0.87
Zhang [9]	0.83	0.81	0.82
Kannan [12]	0.90	0.88	0.89
Sosnovik [22]	0.87	0.85	0.86
Fei [23]	0.89	0.86	0.88

TABLE II: OPTIMIZATION OBJECTIVES OF MODELS

Model	Coverage	Diversity	Redundancy
Niu [5]	0.78	0.72	0.28
Park [6]	0.74	0.70	0.31
Peng [7]	0.76	0.69	0.30
Nagar [8]	0.83	0.80	0.22
Zhang [9]	0.77	0.73	0.29
Kannan [12]	0.85	0.82	0.18
Sosnovik [22]	0.82	0.79	0.21
Fei [23]	0.84	0.81	0.19

B. Optimization Objectives

Optimization objectives measure how models balance coverage, diversity, and redundancy. Table II illustrates that reinforcement learning-based methods (Nagar [8]) achieved high coverage and diversity while maintaining lower redundancy compared to traditional approaches. Kannan et al. [12] further improved diversity with the integration of semantic detectors.

TABLE III: FEATURE COMPLEXITY ACROSS MODELS

Model	Semantic Detectors	Embedding Dim
Niu [5]	10	128
Park [6]	8	64
Peng [7]	12	128
Nagar [8]	6	256
Zhang [9]	5	128
Kannan [12]	25	256
Sosnovik [22]	7	512
Fei [23]	14	256

TABLE IV: CUSTOMIZATION CONSTRAINTS ACROSS MODELS

Model	Max Frames (L)	Max Duration (T)
Niu [5]	120	180
Park [6]	150	240
Peng [7]	100	150
Nagar [8]	90	140
Zhang [9]	110	160
Kannan [12]	80	120
Sosnovik [22]	95	150
Fei [23]	85	130

C.Feature Complexity

The complexity of feature representation significantly impacts summarization accuracy. As seen in Table III, Kannan et al. [12] employed 25 semantic detectors, highlighting their strong reliance on semantic cues. Other

methods varied in the dimensionality of embeddings, ranging from 64 to 512 dimensions, indicating a trade-off between computational cost and representational richness.

D. Customization Constraints

Customization plays a crucial role in aligning summaries with user-defined parameters. Table IV shows maximum frame limits and duration thresholds set by different methods. Park and Cho [6] emphasized longer summaries (150 frames and 240 seconds), while Kannan [12] focused on shorter, more concise summaries.

E. User Interaction Metrics

User interaction is a critical aspect of personalization. As illustrated in Table V, Zhang et al. [9] stand out by incorporating gesture-based controls (12 gestures), while other methods emphasized control actions such as pause or rewind. Explicit use of gaze data was not prevalent among the models listed in this subset.

TABLE V: USER INTERACTION METRICS

Model	Gestures	Gaze Events	Control Actions
Niu [5]	0	0	8
Park [6]	0	0	10
Peng [7]	0	0	7
Nagar [8]	0	0	6
Zhang [9]	12	0	5
Kannan [12]	0	0	9
Sosnovik [22]	0	0	7
Fei [23]	0	0	8

TABLE VI: SCALABILITY AND EFFICIENCY FACTORS

Model	Time Before (s)	Time After (s)	SRR
Niu [5]	60	22	0.58
Park [6]	85	35	0.59
Peng [7]	70	28	0.60
Nagar [8]	55	18	0.67
Zhang [9]	65	24	0.63
Kannan [12]	50	16	0.68
Sosnovik [22]	58	21	0.64
Fei [23]	53	19	0.64

F. Scalability and Efficiency

Scalability and efficiency are key concerns for real-world deployment. Table VI demonstrates that all methods significantly reduced processing times, with Kannan [12] showing the most notable improvement (from 50s to 16s). Similarly, storage reduction ratios (SRR) ranged from 0.58 to 0.68, with the highest observed in semantic and reinforcement learning-based methods.

G. Discussion

The comparative analysis presented in Tables I to VI demonstrates clear distinctions in how different approaches balance efficiency, personalization, and scalability in video summarization tasks. Traditional methods such as those of Niu [5] and Park [6] provide valuable baseline systems that emphasize interactive adaptation and multi-source integration. However, their performance across evaluation metrics such as precision and F1-score shows limitations compared to more advanced deep learning techniques. For example, Kannan et al. [12], who integrated 25 semantic detectors, significantly outperformed others in both recall and precision, highlighting the importance of semantic-aware feature extraction. This suggests that semantic richness directly influences the comprehensiveness of generated summaries. Reinforcement learning–based strategies, as employed by Nagar [8], demonstrated superior optimization of coverage and diversity while effectively minimizing redundancy. This outcome is expected because reinforcement learning directly incorporates reward functions that penalize redundancy and encourage diversity, resulting in more representative summaries. Graph-based models, such as Zhang [9], introduced an interactive and intuitive element through sketch-driven customization, though their quantitative performance in terms of coverage and recall is more modest compared to semantic and reinforcement-based methods.

Emotion-driven and gaze-driven approaches (Ul [15], Katti [17]) emphasize user engagement by focusing on affective and behavioral signals. While these models are not explicitly captured in the current tables, their

reported results suggest a potential to complement semantic or event-based frameworks. A challenge, however, is the acquisition and processing of reliable affective signals at scale, which may limit their deployment in real-world consumer applications. In terms of scalability and efficiency, nearly all approaches showed substantial reductions in processing time and storage requirements. The strongest performer, Kannan [12], demonstrated both the shortest post-summarization processing time and the highest storage reduction ratio (0.68). This indicates that lightweight semantic modeling can simultaneously provide strong performance while remaining computationally feasible, a critical requirement for real-time applications such as mobile devices and streaming platforms.

Taken together, the results highlight a trajectory of progress: from interactive rule-based approaches to semantically enriched deep learning and reinforcement frameworks. The integration of personalization signals, semantic understanding, and computational efficiency appears to define the current state of the art. Nonetheless, challenges remain in balancing personalization with generalization, ensuring robustness to varied video domains, and maintaining efficiency at scale.

Future research in personalized video summarization can progress along several promising directions. One important avenue is multimodal fusion, where audio signals, text transcripts, and motion cues are combined with visual features to yield richer summaries. Another focus lies in emotion and engagement modeling, leveraging affective computing signals such as facial expressions, gaze behavior, and biometric data to enhance personalization. At the architectural level, efforts toward lightweight designs like MobileNets and TinyViTs are essential for enabling efficient real-time summarization on mobile and edge devices. Equally significant is the role of self-supervised learning, which can exploit large-scale unlabeled video data to reduce annotation costs while improving feature generalization. Advanced learning paradigms such as reinforcement and contrastive learning synergy provide further opportunities, balancing diversity and relevance by combining reward-driven policies with embedding-based discrimination. To ensure user trust, explainability and transparency should be prioritized by embedding explainable AI modules that clarify why particular frames or events are selected. Beyond summarization, integration with personalized recommender systems can enable highlight suggestions tailored to user history and contextual preferences. Finally, enhancing cross-domain robustness is critical, with adaptive models designed to handle diverse domains including sports, movies, egocentric videos, and surveillance content with minimal re-training.

V. Conclusion

This review presented an in-depth analysis of personalized video summarization approaches with a focus on deep learning frameworks and mathematical formulations. By examining a wide range of techniques—spanning semantic feature detectors, reinforcement learning, graph-based modeling, emotion-driven methods, and contrastive learning, the study highlighted the evolution from traditional summarization to user-centric, adaptive models. Comparative results demonstrated that semantic-enriched approaches and reinforcement learning strategies consistently outperformed conventional methods in terms of precision, recall, F1-score, and storage reduction. At the same time, graph-based and gaze-driven approaches introduced innovative dimensions of interactivity and engagement, though they often required richer contextual signals. The evaluation further emphasized the importance of balancing accuracy with efficiency. Lightweight architectures such as those employing semantic detectors or contrastive optimization proved effective for real-time applications, especially in mobile and resource-constrained environments. Additionally, the integration of explainable Ai and personalization was identified as a key factor in improving user satisfaction and trust. Overall, the findings indicate that the future of video summarization lies in models that can seamlessly integrate multimodal cues, adapt to diverse user preferences, and maintain computational scalability. Such advancements will ensure that summarization systems not only condense video data effectively but also deliver meaningful and personalized content experiences.

References

[1] P. Saini, K. Kumar, S. Kashid, A. Saini, and A. Negi, “Video summarization using deep learning techniques: a detailed analysis and investigation,” Artificial Intelligence Review, vol. 56, no. 11, pp. 12347–12385, Nov. 2023, doi: 10.1007/S10462-023-10444-0/TABLES/17.

[2] D. M. Argaw et al., “Scaling Up Video Summarization Pretraining with Large Language Models,” Apr. 2024, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/2404.03398v1.

[3] K. Kawamura and J. Rekimoto, “FastPerson: Enhancing Video-Based Learning through Video Summarization that Preserves Linguistic and Visual Contexts,” ACM International Conference Proceeding Series, vol. 1, pp. 205–216, Apr. 2024, doi: 10.1145/3652920.3652922.
[4] R. Wang et al., “Transparency in persuasive technology, immersive technology, and online marketing: Facilitating users’ informed decision making and practical implications,” Computers in Human Behavior, vol. 139, p. 107545, Feb. 2023, doi: 10.1016/J.CHB.2022.107545.
[5] J. Niu, D. Huo, K. Wang, and C. Tong, “Real-time generation of personalized home video summaries on mobile devices,” Neurocomputing, vol. 120, pp. 404–414, Nov. 2013, doi: 10.1016/J.NEUCOM.2012.06.056.
[6] H. S. Park and S. B. Cho, “A personalized summarization of video life-logs from an indoor multi-camera system using a fuzzy rule-based system with domain knowledge,” Information Systems, vol. 36, no. 8, pp. 1124–1134, Dec. 2011, doi: 10.1016/J.IS.2011.04.005.
[7] W. T. Peng et al., “Editing by viewing: Automatic home video summarization by viewing behavior analysis,” IEEE Transactions on Multimedia, vol. 13, no. 3, pp. 539–550, Jun. 2011, doi: 10.1109/TMM.2011.2131638.
[8] P. Nagar, A. Rathore, C. V. Jawahar, and C. Arora, “Generating Personalized Summaries of Day Long Egocentric Videos,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 6832–6845, Jun. 2023, doi: 10.1109/TPAMI.2021.3118077.
[9] Y. Zhang, C. Ma, J. Zhang, D. Zhang, and Y. Liu, “An interactive personalized video summarization based on sketches,” Proceedings - VRCAI 2013: 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry, pp. 249–258, 2013, doi: 10.1145/2534329.2534343.
[10] A. Tejero-De-Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, “Summarization of User-Generated Sports Video by Using Deep Action Recognition Features,” IEEE Transactions on Multimedia, vol. 20, no. 8, pp. 2000–2011, Aug. 2018, doi: 10.1109/TMM.2018.2794265.
[11] W. N. Lie and K. C. Hsu, “Video summarization based on semantic feature analysis and user preference,” Proceedings - IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, pp. 486–491, 2008, doi: 10.1109/SUTC.2008.88.
[12] R. Kannan, G. Ghinea, S. Swaminathan, and S. Kannaiyan, “Improving video summarization based on user preferences,” 2013 4th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2013, 2013, doi: 10.1109/NCVPRIPG.2013.6776187.
[13] B. L. Tseng, C. Y. Lin, and J. R. Smith, “Video personalization and summarization system for usage environment,” Journal of Visual Communication and Image Representation, vol. 15, no. 3, pp. 370–392, Sep. 2004, doi: 10.1016/J.JVCIR.2004.04.011.
[14] Y. Dong et al., “Personalized video summarization with idiom adaptation,” MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia, pp. 1041–1043, Oct. 2019, doi: 10.1145/3343031.3350584.
[15] I. Ul Haq, A. Ullah, K. Muhammad, M. Y. Lee, and S. W. Baik, “Personalized Movie Summarization Using Deep CNN-Assisted Facial Expression Recognition,” Complexity, vol. 2019, no. 1, p. 3581419, Jan. 2019, doi: 10.1155/2019/3581419.
[16] B. Köprü and E. Erzin, “Use of Affective Visual Information for Summarization of Human-Centric Videos,” IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3135–3148, Oct. 2023, doi: 10.1109/TAFFC.2022.3222882.
[17] H. Katti, K. Yadati, M. Kankanhalli, and C. Tat-Seng, “Affective video summarization and story board generation using pupillary dilation and eye gaze,” Proceedings - 2011 IEEE InternationalSymposium on Multimedia, ISM 2011, pp. 319–326, 2011, doi: 10.1109/ISM.2011.57.
[18] G. Mujtaba, A. Malik, and E.-S. Ryu, “LTC-SUM: Lightweight Client-driven Personalized Video Summarization Framework Using 2D CNN,” IEEE Access, vol. 10, pp. 103041–103055, Jan. 2022, doi: 10.1109/ACCESS.2022.3209275.
[19] A. Yoshitaka and K. Sawada, “Personalized video summarization based on behavior of viewer,” 8th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2012r, pp. 661–667, 2012, doi: 10.1109/SITIS.2012.100.
[20] D. R. Olsen and B. Moon, “Video summarization based on user interaction,” EuroITV’11 - Proceedings of the 9th European Interactive TV Conference, pp. 115–122, 2011, doi: 10.1145/2000119.2000142.
[21] M. Fei, W. Jiang, and W. Mao, “Creating personalized video summaries via semantic event detection,” Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 11, pp. 14931–14942, Nov. 2023, doi: 10.1007/S12652-018-0797-0/METRICS.
[22] I. Sosnovik, A. Moskalev, C. Kaandorp, and A. Smeulders, “Learning to Summarize Videos by Contrasting Clips,” Jan. 2023, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/2301.05213v3.
[23] M. Fei, W. Jiang, and W. Mao, “Learning user interest with improved triplet deep ranking and web-image priors for topic-related video summarization,” Expert Systems with Applications, vol. 166, p. 114036, Mar. 2021, doi: 10.1016/J.ESWA.2020.114036.