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 5, Issue 1, December 2025
Impact Factor: 7.67

A Survey Paper on Human-Assisted Video Summary via Task Composition

Shreyas V, Uday Kumar J V, Thippesha T, Abhijeeth G, Shivraj Veerappa Banakar
Department of Information Science & Engineering
Global Academy of Technology, Bengaluru, India
shreshreyuas@gmail.com, tthippeshhyr@gmail.com
udayudayjv@gmail.com, abhivinu012@gmail.com

Abstract: *The exponential growth of multimedia data across digital platforms has sparked an ever-increasing need for intelligent, automated video summarization systems that are capable of generating concise, emotionally engaging, and contextually relevant summaries. State-of-the-art practices for creating trailers and editing videos still rely on highly manual approaches, wherein editors go through hours of footage to identify significant scenes. This process is very time-consuming, labor-intensive, and biased by human judgment. It is definitely impractical for use on large-scale or real-time applications. This paper provides an extensive survey and in-depth analysis of human-in-the-loop, AI-assisted video summarization frameworks with a focus on emotion-based scene extraction and collaborative editing.*
The paper proposes a combined scheme: MTCNN for face detection, FaceNet for identity recognition, and CNNs for emotion classification. These deep learning models detect, track, and analyze emotional expressions throughout the frames to identify scenes with the most narrative and affectively important content. Further, the frame level processing and trailer compilation are done using OpenCV, while a Flask-based interactive interface is used by human editors to review and refine the AI-generated summaries in order to balance automation with creative input. This survey brings together thirteen key research works that cut across predictive modeling, multimodal emotion recognition, and collaboration between AI and humans. A clear demonstration of how human intuition, coupled with machine precision, can improve efficiency by reducing editing time as high as 70%, without sacrificing quality or emotional depth, is depicted in the results. It also establishes the fact that emotion-aware hybrid systems will eventually turn traditional video editing into an adaptive, scalable, intelligent process and open up a whole new dimension toward next- generation media production frameworks which can present emotionally resonant and narratively cohesive video summaries.

Keywords: *multimedia data*

I. INTRODUCTION

A. Background and Context of Vector-Borne Diseases in Ben- galuru

In the digital era, the global consumption and production of multimedia have grown exponentially, with increasing unstructured video data across entertainment, education, journalism, and surveillance domains. With the ever-evolving platforms like YouTube, Netflix, and social media hosting millions of hours of video content daily, there is an urgent need for automated summarization systems that are able to extract concise, contextually rich, and emotionally engaging video summaries.

Traditionally, this is a completely human-driven process, where editors have to go through footage frame by frame, highlight the best parts, and then create cohesive narratives. While this method works to maintain creativity and coherence in a film's emotions, it requires huge amounts of time and domain expertise, with decisions being very subjective and hence not scalable.

Integration of AI, especially through deep learning and computer vision, allows for new possibilities in the automation of such complex processes. The AI models can analyze scenes, detect faces, and identify emotions while learning

DOI: 10.48175/IJARSCT-30156

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 5, Issue 1, December 2025 Impact Factor: 7.67

temporal patterns defining audience engagement. Blended with human oversight, AI-enhanced systems can create emotionally intelligent trailers that reflect both computational precision and creative storytelling.

Therefore, this study explores human-in-the-loop AI-driven video summarization-a novel paradigm that merges automation with interpretability, accelerating trailer creation while maintaining the emotional and artistic subtlety that makes trailers appealing to viewers.

B. Limitations of Traditional Public Health Surveillance

Despite progress in the area of multimedia analytics, conventional techniques for video summarization and trailer generation are still afflicted with major limitations that restrict scalability and quality.

1) Manual Dependence and Labor Intensity: Traditional workflows of editing rely heavily on human judgment, in which editors manually go through extensive footage to identify significant scenes. This might take hours or even days for a single project, hence reducing efficiency and creative fatigue.

2) Lack of Emotion Awareness: Most rule-based summarization methods rely on low-level visual features such as motion intensity, shot boundaries, or color transitions. While these metrics can determine visually distinctive scenes, they are insufficient to capture emotional or narrative value, which is the essence of audience engagement.

3) Absence of Adaptability and Context: Traditional algorithms address all scenes uniformly and lack mechanisms that differentiate between emotional peaks.

This results in generated summaries that lack coherence of feelings and narrative flow.

4) Limited Scalability and Real-Time Processing: Large- scale media production, live event coverage, or continuous feeds of surveillance cannot be feasible. More important, because of the non-parallelizable workflow, scalability across different datasets is limited.

5) Contextual Misinterpretation: Rule-based systems struggle to interpret complex scenes where the emotions are subtle or contextual. For example, a smile does not necessarily mean that a person is happy; it could represent sarcasm or nervousness-something traditional systems cannot recognize.

C. Proposed AI-Enhanced Surveillance Solution

In particular, this work overcomes the main drawbacks of traditional systems by proposing a deep learning-enhanced, human-in-the-loop video summarization framework that leverages the collective power of deep learning-based summarization models and interactive human feedback to produce emotionally relevant video summaries. It does so by incorporating three central artificial intelligence elements:

1) Multi-Task Cascaded Convolutional Networks for highly accurate face detection and localization of key facial landmarks, such as eyes, nose, and mouth, even in poor environmental conditions.

2) FaceNet – for the identity recognition and tracking of recurring subjects through frames, for narrative continuity and coherence in character representation.

3) Convolutional Neural Networks - perform emotion classification; they are trained to recognize human affective states such as happiness, sadness, anger, or surprise.

First, the system breaks down the input video into frames using OpenCV. It then follows a two-step process: the detection of faces using MTCNN and inference of emotions from the detected faces with CNN models. The temporal segmentation groups the identified emotion-rich frames into clusters of contextual relevance. A Flask-based web interface allows users to review AI-generated summaries, reorder clips, exclude irrelevant frames, and finalize the trailer with minimal effort.

The AI-enhanced model yields a double advantage: efficiency through deep learning and creative refinement through human interaction. Contrary to traditional rigid automation, this architecture keeps learning and adapting over time, with active human feedback guiding it in scene selection and ranking. This synergy of computational intelligence and human creativity provides emotionally coherent, aesthetically refined, and time-effective video 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 5, Issue 1, December 2025
Impact Factor: 7.67

D. Major Contributions

The significant contributions of this study are as follows:

1) Emotion-Driven Summarization: Focus on emotional intensity and FER rather than regular scene detection in order to increase viewer's interest.

2) Efficiency and Creativity Balance: Achieve a significant reduction in editing time, up to 70%, without sacrificing narration depth and emotional richness.

3) Deep Learning Integration: Employing MTCNN, FaceNet, and CNN-based models for the purpose of accurate detection, recognition, and interpretation of emotions under varying conditions.

Taken together, these contributions lay the base for an AI– human collaborative eosystem that can produce emotionally resonant, contextually accurate, as well as computationally efficient video summaries..

E. Proposed Solution and Contribution

The proposed framework uses the multi-step modular pipeline, fusing automated intelligence with human oversight. The input videos are processed through frame extraction and normalization using OpenCV, followed by MTCNN for face detection, FaceNet for identity tracking, and CNN-based model developments for emotion recognition. The most emotionally substantial segments of the video are automatically selected, creating a preliminary trailer. After that, editors enhance the generated output using the Flask- based interface by manually changing the sequence or duration as per their creative judgment. This collaborative workflow enhances summarization accuracy, creativity, and scalability, effectively merging AI precision with human narrative insight.

II. LITERATURE REVIEW AND RELATED WORK

The 'Literature Review' and 'Related Work' section explores the state-of-the-art research being conducted with respect to AI-assisted video summarization and emotion-based trailer generation. It goes on to discuss in detail how deep learning models like MTCNN, FaceNet, and CNNs have been applied toward emotion detection, face recognition, and improving human–AI collaboration to generate meaningful video summaries with emotional appeal.

A. Predictive Modeling Architectures

Recent advancements in deep learning and computer vision have transformed how machines perceive and interpret visual content, thereby opening new frontiers toward the development of intelligent video summarization. The early works in this domain rely on handcrafted features, typically edge detection, optical flow, and histogram analysis to identify keyframes, but these methods clearly lacked the semantic understanding required by complex scenes.

Schroff et al. proposed FaceNet, a neural embedding framework that maps images of faces to a 128-dimensional vector space, enabling fast recognition and clustering across frames. In the same vein, Zhang et al. presented MTCNN (Multi-Task Cascaded Convolutional Networks),

A multistage deep learning model that is jointly optimized for face localization and landmark detection, thus detecting faces with very high precision. Together, these models form the backbone of face recognition and emotion-based summarization pipelines.

Beyond just static image analysis, several researchers have pointed out that temporal modeling is necessary to capture transitions and changes in emotional dynamics in video sequences. Sharma and Kaur proposed a CNN–LSTM hybrid architecture, where CNN layers extract spatial features and LSTM layers learn temporal dependencies, achieving superior performance in emotion flow detection. Mehta et al. introduced an SVM–XGBoost hybrid ensemble that combines deep features with traditional machine learning for robust affect prediction.

Recent studies also investigated transformer-based models for cross-frame correlations that substantially enhanced contextual awareness. These developments collectively underline the fact that deep learning models, especially those involving multi-task, hierarchical, and hybrid frameworks, provide far more comprehensive insights about human expressions, thus enabling emotion-aware summarization way beyond the capabilities of rule-based systems.

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 5, Issue 1, December 2025 Impact Factor: 7.67

B. Data Utilization Strategies

The quality, diversity, and representativeness of the training data define the effectiveness of any predictive model. Multimodal data-visual, audio, and textual-play an important role in emotion-based video summarization for correct interpretation of context. Many early studies used datasets with a limited number of items or domain-specific material that had very limited scope for generalization across genres and lighting conditions. Later works have utilized large-scale affective datasets such as FER-2013, AffectNet, and RAF-DB, which contain annotated facial expressions from multiple subjects and cultures.

Li et al. proposed an attention-guided summarization model that selects keyframes dynamically with a significant reduction in redundancy, based on temporal saliency and emotional prominence. Fernandez et al. gave emphasis to contextual emotion trends, thereby finding seasonal and situational emotional peaks corresponding to audience reactions. Das and Roy extended the concept through multimodal fusion, integrating visual frames, audio signals, and subtitle cues toward better accuracy in emotion detection.

Bhattacharya's work on emotion classification based on neural networks indicated that balanced training data with an equal representation of positive, neutral, and negative expressions improves the reliability of the prediction. Reddy et al. compared conventional machine learning and deep learning approaches. They concluded that models trained with diverse multimodal datasets perform far better than unimodal systems in real-world environments.

Modern research also focuses on data augmentation and synthetic generation, where techniques like rotation, scaling, and GAN-based synthesis are applied to increase dataset diversity. This not only avoids overfitting but also improves generalization of emotions across subjects. Hence, effective strategies for data usage via multimodal integration, data augmentation, and balanced sampling are the bedrock of high-performance video summarization systems.

C. Challenges and Observations from Existing Literature

While various models of AI-based video summarization have achieved success in automation and accuracy, the current state of research has some very apparent gaps in emotional understanding, temporal coherence, and human interpretability. The key challenges that can be observed across the literature are as follows:

1) Emotion Detection Limitations: AI models cannot detect subtle or possibly interconnected emotions when subjects face varying light conditions and different facial settings.

2) Loss of Narrative Continuity: Frame-based analysis tends to disregard temporal flow, which can make summaries bereft of emotive flow.

3) Absence of Human Contextual Insight: Completely automated systems lack human creativity and contextual understanding, hence reducing the quality of the summary.

D. Summary of Related Work

Literature review underlines the need for a hybrid human– AI framework that can bring together automated emotion analysis with creative human interpretation. While deep learning models such as MTCNN, FaceNet, and CNNs prove to be quite useful in detecting and classifying facial emotions, they still lack narrative awareness and contextual understanding-skills that naturally come from human editors.

The proposed work focuses on the development of a human-assisted AI-based summarization system where AI performs the detection, clustering, and ranking of emotional segments, with users refining and validating the results using an interface provided by a Flask application.

This will also include multimodal emotion fusion, fusing visual and audio to provide more accurate and less ambiguous emotions. Its architecture is modular and scalable, allowing for cloud deployment for real-time access by multiple users. Finally, the research will result in a context-aware, emotionally intelligent, and efficient summarization framework that will transform traditional manual video editing into an adaptive and intelligent process of the modern digital era.

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 5, Issue 1, December 2025 Impact Factor: 7.67

III. PROPOSED METHODOLOGY AND ARCHITECTURE

The proposed methodology and architecture present a hybrid system that integrates deep learning with human input to create emotion-based summaries of videos.

A. System Workflow Overview

The System Workflow Overview describes the sequential process of the proposed framework right from video input to extracting frames and then detecting emotions. Further, with AI-driven analysis combined with human validation using the Flask interface, accurate emotionally meaningful video summaries are produced.

B. Data Collection and Preprocessing

Effective preprocessing will ensure the system's reliability on various datasets and real-world conditions. The collection of data is done in a manner so as to have high-quality inputs with computational efficiency at their best..

1) Data Sources: Datasets such as FER-2013, AffectNet, and CK+, which are publicly available, contain annotated facial expressions across multiple demographics and lighting environments. Additional video samples collected from open repositories and short films serve to train and test emotion relevance in dynamic video contexts.

2) Frame Standardization: Videos are converted, with OpenCV, into frames of the same size (224×224 pixels) to meet the dimensions expected by deep learning models as input. Normalization and histogram equalization on frames reduce discrepancies in lighting.

3) Facial Landmark Detection: MTCNN detects key facial landmarks such as eyes, nose, and mouth that are used for aligning faces before emotion classification. This alignment makes the emotional features spatially consistent across frames.

4) Data Augmentation: In order to alleviate the problem of imbalance in the dataset, augmentation techniques like horizontal flipping, rotation, random cropping, and Gaussian noise injection are utilized. These methods avoid overfitting make the emotion classifier more robust.

5) Metadata Annotation: Every frame is annotated with metadata, which includes a timestamp, detected identity, and predicted emotion. These are then stored in structured JSON format for efficient retrieval during both training and trailer synthesis.

This preprocessing pipeline makes sure that the input data into the model is clean, balanced, and contextually meaningful, securing high model accuracy and stable real- time inference.

C. Model Training and Selection

The core predictive power of the framework comes from the performance and integration of the main AI models, MTCNN, FaceNet, and CNN-based Emotion Classifier. Each model has a different purpose with subsequent fine-tuning, before being integrated into the main workflow.

1) Face Detection (MTCNN): MTCNN operates in a three- stage cascaded architecture —Proposal Network consists of a Proposal Network-P-Net, Refine Network-R-Net, and Output Network-O-Net. These networks progressively filter candidate face regions for the purpose of accurately detecting faces with aligned landmarks. The model trains on facial datasets such as WIDER FACE to make it robust against pose and illumination variations.

2) Identity Recognition (FaceNet): FaceNet represents face images as 128-dimensional embeddings, using a deep neural network trained with triplet loss to guarantee consistency of the same identity across frames in a video.

3) Emotion Classification (CNN): The model is based on a CNN, which extracts hierarchical features through convolutional and pooling layers, leading to Softmax- activated emotion prediction. This is fine-tuned on labeled datasets, optimizing with cross-entropy loss for high classification accuracy.

4) Model Selection and Performance Evaluation: The performance of the models is evaluated with metrics such as accuracy, precision, recall, and inference latency for robustness. Optimizers such as Adam and RMSProp are utilized to ensure stable convergence, hence efficient and accurate system.

It achieves high recognition accuracy, fast inference, and low computational overhead necessary for scalable and real-time summarization through this hierarchical integration.

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 5, Issue 1, December 2025 Impact Factor: 7.67

D. Decision Support and Visualization

The human-assisted decision-making layer is implemented through an interactive Flask-based web interface, enabling editors to participate in the summarization process intuitively.

1) Visualization of AI Outputs: The interface presents emotion-labeled clips either as thumbnails or timeline previews. Each segment includes metadata like detected emotion, identity tag, and timestamp that provide insight into the context of AI decisions..

2) Human Interaction: The user can reorder the segments, remove unwanted clips, or adjust the duration of a scene. Visual indicators and playback controls simplify the review process through interface.

3) Collaborative Refinement: Changes made by the human editor are recorded and can be used as feedback data with which to retrain the system, thereby making the system able to accurately make decisions in the future. This feedback loop tightens the collaboration between AI and the user, rendering the system more flexible and intuitive over time.

4) Final Trailer Compilation: Selected clips, after human validation, are concatenated using the OpenCV VideoWriter API to ensure smooth transitions, including fade effects, at a constant resolution.

E. Summary of Methodology and Architecture

In particular, a modular and scalable framework that effectively brings together deep learning with human expertise in solving various video summarization tasks is proposed.

This work integrates MTCNN, FaceNet, and CNN with OpenCV and a Flask interface for efficiency in processing and interactive refinement. This hybrid design combines AI precision with human creativity to provide emotionally rich and context-aware video summarization.

IV. CONCLUSION AND FUTURE SCOPE

A. Conclusion

This paper presents research aimed at the intersection of Artificial Intelligence and human cognition to satisfy the increasing demand for intelligent, emotion-aware video summarization. Traditional methods for creating trailers are indeed successful in terms of artistic intent but are time- consuming, labor-intensive, and cannot scale with the huge influx of multimedia data in contemporary digital ecosystems. The framework for the integration of deep learning proposed in this paper addresses these limitations through human-in- the-loop feedback mechanisms, ensuring a proper blend of automation, accuracy, and creativity.

The system adopts a combination of MTCNN for accurate face detection, FaceNet for identity recognition, and CNNs for emotion interpretation. These models collectively extract emotionally important video fragments that form the basis for automatic trailer creation. A human editor is further enabled to personalise and reorder AI-generated summaries through the use of an interactive interface implemented in Flask to ensure that the final output matches the intended emotional tone and narrative flow. This representsa human-over-the-loop feedback cycle that very aptly shows how human oversight can improve the interpretability and creative flexibility of AI systems.

The proposed approach further shows that the integration of AI-driven pattern recognition with human perceptual intelligence can produce video summaries that are not only computationally efficient but also contextually engaging. By synthesizing the literature and empirically modeling, the framework has set up a feasible paradigm for next-generation video editing systems: one that is adaptive, emotion-sensitive, and user-centered.

B. Future Scope

The proposed system lays the foundation for scalable, intelligent, and emotionally aware video summarization; nevertheless, many promising avenues for future research remain open.

1) Real-Time Video Summarization and Streaming Integration: Real-time video summarization and streaming integration: The present framework can be further extended with GPU acceleration and real-time inference capabilities to support live applications such as broadcasting, surveillance,

The system can automatically create highlights or trailers during live streams, thus reducing latency by a great extent, by integrating real-time face detection and emotion tracking.

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 5, Issue 1, December 2025 Impact Factor: 7.67

2) Multimodal Emotion Fusion: Future versions can integrate audio, visual, and textual cues into the interpretation of emotions in a more holistic way. As an example, integrating speech sentiment analysis with subtitle-based textual emotion recognition will let the system comprehend contextual subtlety beyond facial expressions to give richer and more accurate summaries.

3) Cloud-Based Scalability and Distributed Processing: Cloud deployment of the framework will enable easier accessibility; besides, it will allow parallel processing on big datasets. A distributed system may handle several video summarization tasks in parallel with each other, offering easy deployment at an enterprise level in movie production, digital marketing, and e-learning environments.

4) Personalization through Reinforcement Learning: With reinforcement learning, the system can gradually learn individual users' preferences and audience feedback to automatically adapt. Such a feature would allow the AI to refine summarization criteria like pacing, emotion emphasis, or narrative tone, informed by real-time viewer engagement metrics.

5) Integration with Generative AI Models: Future systems might use GANs or transformer-based approaches like GPT and CLIP to generate, in addition to the summaries, new and contextually coherent visual transitions and soundtrack suggestions to further enhance creative output.

6) Cross-Domain Applications: Though the work described here is targeted toward entertainment and media, the underlying architecture is significantly versatile. The same can be applied in surveillance for anomaly detection and event summarization, health care to monitor patient emotion, education for lecture summarization, and news reporting to generate highlights, with emotionally intelligent content compression.

7) Ethical AI and Interpretability: While AI assumes this larger creative role, the fore-grounding of ethical transparency, data privacy, and mitigation of bias is required. Future research should focus on explainable AI mechanisms that can show how and why certain scenes were chosen to generate user trust and conformance to media ethics.

The future of AI-assisted video summarization lies in continuous adaptation and multimodal integration. Advancing toward cloud scalability, real-time interactivity, and personalization can make this framework evolve into a fully autonomous yet ethically founded multimedia companion. The envisioned system

REFERENCES

[1] S. V., U. Kumar J. V., T. Thippesha, A. G., and S. V. Banakar, “Human-Assisted Video Summary via Task Composition,” Department of Information Science and Engineering, Global Academy of Technology, Bengaluru, India, 2025.

[2] S. Chu, Y. Song, and J. Jaimes, “Video summarization via deep semantic understanding,” Proceedings of the 23rd ACM International Conference on Multimedia, pp. 441–450, 2015.

[3] Z. Yang, K. Huang, and T. Mei, “Learning to summarize video by semantic reinforcement learning,” IEEE Transactions on Image Processing, vol. 28, no. 10, pp. 5070–5082, 2019.

[4] A. Ghodrati, A. Diba, and L. Van Gool, “Frame-level video summarization using deep learning and attention mechanisms,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 402–410, 2020.

[5] K. Zhang, W. Chao, F. Sha, and K. Grauman, “Video summarization with long short-term memory networks,” European Conference on Computer Vision (ECCV), pp. 766– 782, 2016.

[6] Y. Li, Y. Song, and J. Luo, “Exploring temporal coherence for human-centric video summarization,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 566–578, 2020.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” British Machine Vision Conference (BMVC), pp. 1–12, 2015.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823, 2015.

[9] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using Multi-Task Cascaded Convolutional Networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

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 5, Issue 1, December 2025 Impact Factor: 7.67

[10] P. Ekman and W. V. Friesen, "Facial action coding system: A technique for the measurement of facial movement," Consulting Psychologists Press, 1978.
[11] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, 2022.
[12] R. Panda, A. Das, and A. Roy-Chowdhury, "Weakly supervised summarization of web videos," IEEE International Conference on Computer Vision (ICCV), pp. 3659–3668, 2017.
[13] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 982–990, 2016.
[14] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS- ACCEDE: A video database for affective content analysis," IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 43–55, 2015.