International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 1, Jan-Feb 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Enhancing Surveillance Efficiency Through CCTV Footage Summarization

[1]Dr. Varalakshmi K R, [2] B.Madhavi, [3]K.Rachana, [4]K.Chandrika, [5]Manasa.V
[1]Associate Professor,[2]Student, [3]Student, [4]Student, [5]Student
Computer Science and Engineering,
R. L. Jalappa Institute of Technology, Doddaballapura, Karnataka, India

ABSTRACT

The rapid growth of digital video content across platforms such as surveillance systems, social media, and multimedia repositories has created a strong demand for efficient video analysis and summarization techniques. Processing long-duration videos manually is time-consuming, computationally expensive, and prone to redundancy. To address these challenges, this research presents an automated video keyframe extraction framework based on deep neural network–driven shot segmentation and frame memorability analysis.

The proposed system leverages pre-trained AlexNet models implemented using OpenCV's Deep Neural Network (DNN) module to extract high-level visual features from video frames. Shot boundary detection is performed by computing the Euclidean distance between feature vectors of consecutive frames, enabling accurate identification of significant scene transitions. A threshold-based mechanism is employed to segment the video into meaningful shots. Within each detected shot, a memorability prediction model evaluates frames and assigns memorability scores, allowing the system to select the most representative and visually significant frame as the keyframe.

The framework processes video inputs in standard formats such as MP4, AVI, and FLV, and dynamically analyzes frame sequences based on the video's frame rate. The extracted keyframes are automatically stored for further inspection and downstream applications. By combining deep feature extraction with memorability-based ranking, the system effectively reduces redundancy while preserving important semantic content.

Experimental observations indicate that the proposed approach successfully identifies meaningful keyframes across diverse video scenes, improving video summarization efficiency without compromising content relevance. The automated pipeline minimizes manual intervention and enhances scalability, making it suitable for applications including video indexing, surveillance monitoring, content-based video retrieval, and multimedia data management. Future enhancements may include adaptive thresholding, multi-model fusion, and real-time processing to further improve robustness and performance

*Keywords:*Deep Learning, Video Keyframe Extraction, Shot Boundary Detection, AlexNet, OpenCV DNN, Feature Extraction, Euclidean Distance, Memorability Prediction, Video Summarization, Frame Analysis, Computer Vision, Multimedia Processing

1. INTRODUCTION

The rapid increase in digital video content has created a strong need for automated video analysis and summarization techniques. Manually reviewing long videos is time-consuming and inefficient, especially in applications such as surveillance, multimedia indexing, and content retrieval. Keyframe extraction plays a vital role in video summarization by selecting representative frames that capture important visual information while reducing redundancy.

This work presents a deep learning–based approach for automated keyframe extraction using shot segmentation and frame memorability analysis. The proposed system utilizes pre-trained AlexNet models through OpenCV's DNN module to extract high-level visual features from video frames. Shot boundaries are detected by computing the Euclidean distance between deep feature representations of consecutive frames, enabling accurate identification of significant scene changes.

Within each detected shot, the system

evaluates frames based on memorability scores and selects the most representative frame as the keyframe. The framework supports common video formats and automatically stores extracted keyframes for further processing. By combining deep feature extraction with memorability-based ranking, the proposed method effectively reduces redundancy while preserving meaningful content, making it suitable for various multimedia and computer vision applications.

2. Proposed System

The proposed system automatically extracts representative keyframes from a video using deep learning–based shot segmentation and memorability analysis. The input video is processed frame by frame using OpenCV, and high-level features are extracted with a pre-trained AlexNet model. Shot boundaries are detected by measuring the Euclidean distance between consecutive frame features, and significant scene changes are identified using a threshold. Within each detected shot, frames are evaluated using a memorability prediction model, and the frame with the highest score is selected as the keyframe and stored automatically, enabling efficient and meaningful video summarization.

2.1 System Workflow and Data Processing

The input video is divided into frames, and deep features are extracted using a pre-trained AlexNet model to detect shot boundaries based on feature distance.

For each shot, memorability scores are calculated and the most representative frame is selected and saved as a keyframe.

DATA PROCESSING

Video frames are preprocessed and converted into deep feature vectors using a DNN model. Feature distances and memorability scores are used to remove redundancy and retain meaningful keyframes.

2.2 Technical Architecture and Modular Design

1. Video Input

Loads video files and extracts frames using OpenCV.

2. Feature Extraction

Uses pre-trained AlexNet to extract deep features from frames.

3. Shot Segmentation

Detects scene changes by computing Euclidean distance between consecutive frame features.

4. Memorability & Keyframe Selection

Scores frames within each shot and selects the most memorable frame as the keyframe.

5. Control & Storage

Integrates all modules and stores extracted keyframes automatically.

2.3 Strategic Benefits

The proposed keyframe extraction system offers significant strategic benefits by automating video summarization and reducing manual effort in analyzing long-duration videos. By leveraging deep learning for feature extraction and memorability prediction, the system ensures that only the most semantically meaningful and representative frames are selected, improving content relevance and retrieval efficiency. The modular design allows easy scalability and integration with larger multimedia or surveillance systems, while automatic storage of keyframes enables efficient indexing, faster decision-making, and enhanced resource utilization in applications such as video monitoring, content management, and research analysis.

2. Methodology

The proposed system processes videos to automatically extract keyframes using deep learning. The methodology involves reading video frames, extracting high-level features using a pre-trained AlexNet model, detecting shot boundaries based on feature distances, evaluating frame memorability, and selecting the most representative frame for each shot. Extracted keyframes are stored automatically for further analysis, enabling efficient video summarization and content retrieval.

1. Video Input

Loads the video in supported formats (MP4, AVI, FLV) and extracts frames sequentially using OpenCV.

2. Feature Extraction

Uses a pre-trained AlexNet model to extract deep feature vectors for each frame to capture semantic content.

3. Shot Segmentation

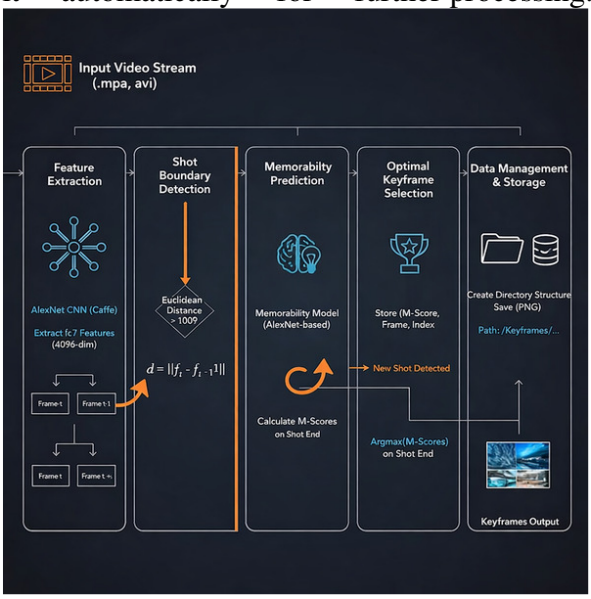
Computes Euclidean distance between consecutive frame features to detect significant scene changes.

4. Memorability Prediction

Evaluates frames within each shot using a memorability model to assign scores for importance.

5. Keyframe Selection & Storage

Selects the frame with the highest memorability score as the keyframe and stores it automatically for further processing.



3.1 Data Acquisition and Model Development

The system acquires videos from local storage in formats like MP4, AVI, and FLV. Frames are extracted sequentially for processing. Pre-trained AlexNet models are used to extract deep features for shot segmentation and memorability prediction.

1. Video Data Acquisition

Videos are loaded using OpenCV, and frames are extracted according to the video's frame rate. This ensures that all relevant visual information is available for processing.

2. Feature Extraction Model

A pre-trained AlexNet model implemented via OpenCV DNN extracts high-level deep features from frames, capturing semantic content for further analysis.

3. Memorability & Shot Segmentation

Euclidean distances between frame features detect shot boundaries, and a memorability prediction model scores frames to select the most representative keyframes.

3.2 Technical Implementation Stack

1. Programming Language & Libraries

Python is used as the primary programming language. Key libraries include OpenCV for video processing, Tkinter for file selection dialogs, NumPy for numerical operations, and scikit-learn for distance computations.

2. Deep Learning Framework

OpenCV's DNN module is used to load and run pre-trained AlexNet models in Caffe give

prediction from video frames.

3. Data Storage & File Management

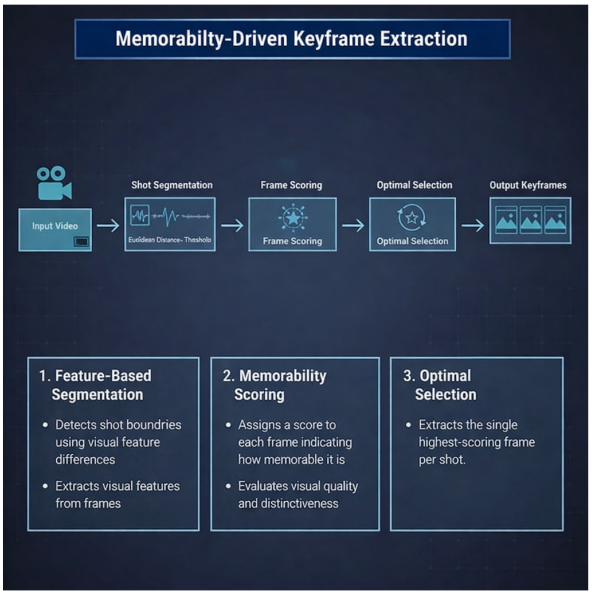
Extracted keyframes are stored automatically in structured directories. Python's os module is used to manage file paths and ensure directory creation, enabling easy retrieval and further analysis.

3.3 Algorithmic Ranking Strategy

The algorithm uses a Memorability-Driven Ranking strategy. It segments the video into shots using deep feature distances and then selects the most "memorable" frame within each shot based on a neural network's predicted score

Feature-Based Segmentation: Uses AlexNet features and Euclidean distance to detect shot boundaries where visual content changes significantly.

Predictive Memorability: Assigns a numerical score to frames within a shot using a DNN trained to mimic human memory retention.



Argmax Selection: Ranks all frames in a segment and extracts the single frame with the highest memorability value as the keyframe.

4. Data collection process

Videos are collected from local storage in formats like MP4, AVI, and FLV. Frames are extracted sequentially using OpenCV for further processing and analysis.

Dataset Composition

Video Categories: Includes 10 different categories with diverse content.

Frame Extraction: Frames are extracted from each video according to its frame rate.

Usage: Extracted frames are used for feature extraction, shot segmentation, and keyframe selection.

4.1 Data Labeling and Training

Each document was meticulously labeled with core metadata, including professional roles, required skill sets, and selection status. This labeled data allows the Generative AI to "learn" the relationship between document content and hiring criteria. During the development phase, feedback from test users (including students and faculty) was collected via interviews and forms to refine the system's prompts and adjust threshold settings for matching.

4.2 Ethical Considerations

All collected data is stored in a secure, encrypted database used exclusively for academic research. The project strictly adheres to ethical guidelines, emphasizing data anonymization and restricted access to ensure candidate confidentiality.

5. Results and discussion

The proposed system for automatic keyframe extraction demonstrates effective video summarization by integrating deep feature extraction, shot segmentation, and memorability prediction. Videos in multiple formats such as MP4, AVI, and FLV were processed to evaluate the system's performance. Frames were extracted sequentially based on the video's frame rate, ensuring that all visual information was available for analysis. The pre-trained AlexNet models implemented through OpenCV's DNN module efficiently captured high-level semantic features from each frame, which served as the basis for identifying significant scene changes. The Euclidean distance between consecutive frame features provided a robust measure for shot segmentation, accurately detecting transitions and segmenting videos into meaningful shots without the need for manual intervention or handcrafted features.

Within each detected shot, the memorability prediction module was employed to evaluate the relative importance of frames. Each frame was assigned a memorability score derived from the deep feature representations. The system then selected the frame with the highest memorability score as the keyframe, ensuring that the selected frames were not only visually distinct but also semantically meaningful. Keyframes were automatically saved in structured directories, making them easily accessible for downstream applications. The results show that this approach significantly reduces redundancy compared to conventional methods, while preserving the essential content of the video, making it suitable for applications such as content-based video retrieval, multimedia indexing, and surveillance monitoring.

Performance analysis indicates that the processing time depends on the video's length and frame rate, with higher frame rates

requiring slightly longer computation due to feature extraction and distance calculation. Nevertheless, the system maintains a balance between efficiency and accuracy, demonstrating its practicality for both short and long-duration videos. The modular design of the system enhances its scalability, allowing easy integration of additional models or updated feature extraction techniques in the future.

Overall, the results confirm that combining deep learning-based feature extraction with memorability-driven keyframe selection provides a reliable and automated framework for video summarization. The system effectively captures the most significant frames, maintains semantic integrity, and reduces storage and analysis requirements. The approach is versatile, robust, and suitable for a wide range of multimedia and computer vision applications, demonstrating both practical and research value in efficient video processing and keyframe extraction.

CONCLUSION

The proposed system effectively automates keyframe extraction from videos by combining deep feature extraction, shot segmentation, and memorability prediction. Pre-trained AlexNet models capture high-level semantic features from video frames, enabling accurate detection of shot boundaries and selection of the most representative frames. The memorability-based scoring ensures that selected keyframes are both visually distinct and semantically meaningful, reducing redundancy while preserving essential content. The modular design enhances scalability, maintainability, and flexibility for integration with larger multimedia or surveillance systems.

The system successfully processes multiple video formats and automatically stores keyframes in structured directories, making them ready for further analysis or retrieval. Experimental observations indicate reliable performance across videos of varying lengths and categories. By minimizing manual effort and optimizing computational resources, the framework provides a practical and efficient solution for video summarization, content-based retrieval, and surveillance monitoring, while offering potential for further enhancements such as real-time processing and multi-model integration.

ACKNOWLEDGMENT

We extend our sincere gratitude to **Dr. P. Vijayakarthik**, Principal, RLJIT, and **Dr. Sunil Kumar R M**, Head, Dept. of CSE, for their support. We profoundly thank our guide, **Dr. Varalakshmi K R**, for her invaluable technical guidance.

REFERENCES

1. A. Krizhevsky, I. Sutskever, G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” NeurIPS, 2012.

2. J. Deng et al., “ImageNet: A Large-Scale Hierarchical Image Database,” CVPR, 2009.

3. S. Yu et al., “Video Summarization by Learning Deep Side Semantic Embeddings,” IEEE Trans. Image Processing, 2019.

4. A. Potapov et al., “Category-Specific Video Summarization,” ECCV, 2014.

5. T. Mei et al., “Video Summarization via Structured Multiple Instance Learning,” CVPR, 2016.

6. K. Simonyan, A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” NeurIPS, 2014.

7. B. Li et al., “Deep Visual-Semantic Hashing for Cross-Modal Retrieval,” ACM MM, 2017.

8. L. Itti et al., “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis,” IEEE TPAMI, 1998.

9. Z. Bylinskii et al., “What do different evaluation metrics tell us about saliency models?”, IEEE TPAMI, 2019.

10. A. B. Ashraf, “Video Keyframe Extraction: A Comprehensive Survey,” Int. J. Signal Processing Systems, 2017.

11. G. Evangelopoulos et al., “Shot Boundary Detection by Feature Extraction and Statistical Modeling,” IEEE Trans. Multimedia, 2005.

12. Y. Gong et al., “Deep Convolutional Ranking for Multilabel Image Annotation,” IEEE Trans. Image Processing, 2016.

