Static Video Summarization Using Transfer Learning and Clustering

Shamal Kashid[1(✉)], Lalit K. Awasthi[1], Krishan Berwal[2], and Parul Saini[1]

[1] Computer Science and Engineering, NIT Uttarakhand, Srinagar, India
{kashid.shamalphd2021,parulsaini.phd2020}@nituk.ac.in
[2] Faculty of Communication Engineering, Military College of Telecommunication Engineering, Indore, India
k2b@ieee.org

Abstract. This research presents a novel approach to video summarization (VS) for efficiently extracting keyframes, reducing video redundancy while preserving essential content. The method combines deep learning and clustering techniques to generate concise video summaries by analyzing videos frame by frame. Using a dual convolutional neural network (CNN), we extract deep-level features. K-means clustering is then applied to group the feature descriptors of the video frames into keyframes and non-keyframes. This K-means clustering-based VS (KVS) method effectively selects the most relevant frames from the extracted features. Our proposed KVS approach outperforms existing VS techniques, achieving an average F-score of 73.1 and 76.3 on two benchmark datasets, Open Video and YouTube, respectively, demonstrating its superior ability to produce informative video summaries.

Keywords: Convolutional Neural Network · Clustering · Transfer Learning · Video Summarization

1 Introduction

The rapid growth of digital video content across various platforms necessitates the development of efficient methods for extracting and presenting relevant information concisely. VS, a critical area within multimedia content analysis, addresses this need by generating brief, informative representations of video content that retain essential information. These summaries enhance quick content browsing and retrieval, as well as the efficiency of tasks like video indexing and recommendation systems. Recent advancements in artificial intelligence and deep learning have significantly impacted VS techniques, enabling both supervised and unsupervised approaches [1]. Supervised methods leverage annotated datasets to learn patterns and predict keyframes or segments that best represent the video's content. In contrast, unsupervised techniques, such as clustering algorithms like K-means, autonomously identify representative frames based on visual similarities. These methods, often enhanced by deep learning architectures, have shown considerable promise in producing accurate and contextually

S. Tiwari et al. (Eds.): AI4S 2024, CCIS 2243, pp. 68–76, 2025.
https://doi.org/10.1007/978-3-031-81369-6_6

relevant video summaries, applicable across various domains, from consumer video platforms to professional surveillance systems [2].

The exponential development of multimedia content on the Internet driven by the rise of inexpensive electronic devices equipped with developed capabilities, presents significant challenges to networks due to the continued increase of extensive video data [3,4]. To mitigate these challenges, researchers are actively exploring solutions focused on efficient video data handling. VS is a critical technique in video analysis, condensing lengthy videos into concise yet informative representations while preserving their core semantics. This method generates succinct abstracts that authorize users to grasp the video's important content quickly [5]. This approach not only enhances accessibility but also optimizes bandwidth utilization and faster content retrieval in multimedia environments.

Researchers have extensively studied VS, broadly categorizing it into two types: static and dynamic VS [9]. Static VS typically consists of a storyboard with keyframes that represent the video's content, primarily focusing on visual data while omitting audio messages. In contrast, dynamic VS creates a video clip that integrates visual data, text data, and audio data. While dynamic VS provides a more comprehensive representation, static VS is simpler to navigate and helps reduce the computational complexity associated with video retrieval and analysis [3,4]. In VS, the problem can be viewed as a clustering issue, where the selected keyframes represent the entirety of the video.

This research presents a novel technique, KVS, for developing summaries of videos across various categories. It addresses the limitations of recent methods and emphasizes the use of deep learning to identify meaningful frames in videos. The proposed work makes key contributions in the following areas:

1. Development of a VS system capable of producing effective summaries for videos from various categories.
2. Utilization of deep frame features extraction using the transfer learning technique of Dual-CNNs with the help of GoogleNet and ResNet-50.
3. We present a method to summarize social media videos using the Dual-CNN model and K-means clustering.

The paper organizes the remaining part as follows: Sect. 2 represents existing video summarization techniques. Section 3 discusses the proposed VS model in detail. Section 4 discusses the VS analysis details. Section 5 discusses the experimental results and performance analysis of the proposed framework. Section 6 provides the article's conclusion and future directions.

2 Related Work

Video summarization techniques concentrate on extracting significant keyframes from videos. A variety of methods have been employed, including feature-based techniques that analyze components such as color, motion, objects, gestures, speech, and audio-visual cues. Additionally, clustering-based approaches, such as k-means clustering, partitioning, and spectral clustering, have been utilized.

Other methodologies include shot selection-based VS, event-based VS, and trajectory-based VS techniques [9].

In a VS, proposed method selects keyframes by the combination of memorability and entropy scores. This methodology checks how easy it is to remember frames along with their information entropy to see how important they are in the summarization process [18]. The cost-effective VS method presented in [19] integrates aesthetic features with deep CNNs to create video summaries. Furthermore, Muhammad et al. [20] propose an architecture specifically designed for summarizing surveillance videos, addressing the challenges posed by devices with limited resources and low complexity. This approach effectively operates in resource-constrained environments while maintaining robust VS capabilities.

The rise of deep learning (DL) has spurred significant research into DL-based VS, categorized into supervised or unsupervised methods based on the DL algorithms used [12]. One instance is equal frame partition-based VS, which applies an equal partition-based clustering technique to group the entire video into keyframes. Another approach, ESVS (Eratosthenes Sieve-based VS) [14], focuses on delivering concise and intelligent video abstraction, often referred to as event summarization.

The proposed methodology aims to improve unsupervised static VS by integrating spatial characteristics from deep neural networks. We achieve this by simplifying the summarization process through efficient feature extraction and clustering methods. The approach also incorporates evaluation metrics on benchmark datasets to demonstrate the model's effectiveness. Previous techniques often focused on spatial or temporal characteristics, ignoring a comprehensive understanding. This approach is particularly useful for dealing with vast video datasets.

3 Proposed Model

This paper presents a novel technique for generating static video summarization that produces high-quality summaries across a diverse range of video categories. The KVS method extracts each frames from the input video, eliminates redundant frames, and processes them using CNNs to extract feature vectors. K-means clustering then classifies these feature vectors into keyframes and non-keyframes. The method aims to preserve the original video's captivating elements. Figure 1 illustrates an detailed steps of the proposed KVS approach.

3.1 Frames and Feature Extraction

We preprocess the videos by converting them into individual frames, extracting features from these frames, and mapping the extracted features. This process maintains the original quality of the videos without applying grayscale conversion or pre-sampling techniques. We use CNN, a deep learning technique, for feature extraction from preprocessed video frames. We employ pre-trained CNN models because comprehensive datasets for training are not easily accessible. We

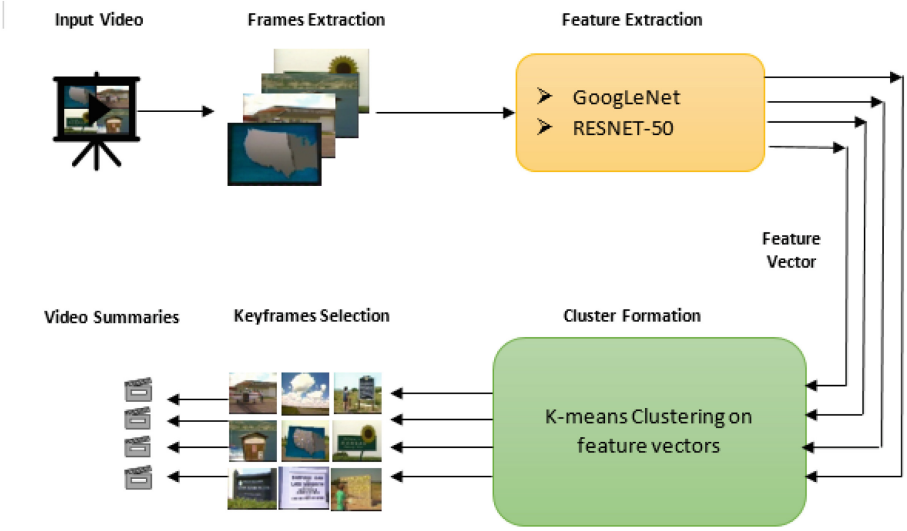

Fig. 1. Proposed KVS Model

preprocess the dataset before feeding it to the model. This includes frames feature extraction with the help of the GoogLeNet [17] and ResNet-50 [16] models together. These networks use the ImageNet dataset, which contains millions of image data, as their training dataset. The feature extraction phase emphasizes the fully connected layer labeled as 'predictions.' A fusion approach combines frames features extracted from both the ResNet-50 and GoogLeNet architectures. Feature fusion combines the features extracted from both networks. These fused features can capture broader information from the input frames.

3.2 K-Means Clustering

This approach leverages the K-means clustering algorithm, an unsupervised iterative learning method, to perform video summarization. The algorithm partitions video frames into clusters based on their visual similarities, utilizing euclidean distance as the measure of similarity. Each cluster represents a group of frames that are similar to each other, while frames in different clusters are distinct. To effectively summarize the video, the algorithm selects representative frames from each cluster to identify keyframes. The silhouette score (SS) determines the optimal number of clusters, which significantly influences the summarization's quality. The silhouette score determines the degree of similarity between a frame and its own cluster, indicating the appropriateness of the frame clustering. By maximizing the silhouette score, we ensure that the frames within each cluster are highly similar and distinct from frames in other clusters, leading to a more accurate and meaningful video summary [8].

$$SilhouetteScore = \frac{a-b}{max(a,b)} \tag{1}$$

The silhouette coefficient, ranging from –1 to 1, measures cluster quality, with values close to 1 indicating well-separated clusters. To determine the optimal number of clusters for each video, we select the SS that is closest to 1. We calculate the SS based on the average distance between a frame and other frames within the same cluster, ensuring that the selected clusters are both cohesive and well-separated.

3.3 Keyframe Selection

This step is designed to identify the frames that most accurately represent the content of each cluster. Initially, we determine the centroid frame for each cluster. Subsequently, we compute the dissimilarity between the centroid and all other frames within that cluster. The frame exhibiting the highest dissimilarity is then selected as the representative for that cluster. Finally, the selected keyframes are aggregated to produce the video summary, effectively conveying the core aspects of the VS process.

4 Performance Analysis

This section uses VS analysis to examine the effectiveness of the proposed KVS model.

VS Analysis: Examining the video summary depends on individual interests and opinions. We visually analyze keyframes for qualitative analysis and compare them with the ground truth summaries. Ground truth summaries are the ideal expected summary provided by humans or users. For quantitative analysis, we utilize three performance metrics:P (precision), R (recall), and F-measure metrics.The marked reference summary is used as a frame unit to estimate the P, R, and F-measures [7]. The harmonic mean of P and R represents the F-measure. It is calculated as $F - Measure = \frac{2.P.R}{(P+R)}$ where P and R are calculated as $P = \frac{A \cap B}{B}, R = \frac{A \cap B}{A}$.

A is the number of frames in the ground truth summary, and B is the number of frames in the proposed summary. A higher F-measure value suggests a more precise approach. To determine P and R, we compare the keyframes in the proposed summary with those in the reference summary. We visually examine the keyframes and compare them with the ground truth summaries for qualitative analysis.

Datasets Used: We evaluate the proposed KVS using two diverse datasets: Open Video (OV) [6] and YouTube (YT) [6]. The OV dataset comprises 50 videos with a frame size of 352×240 resolution. Each video of OV dataset is in the for of MPEG-1 format, sound with color. The OV dataset's total size is 763 MB, encompassing various genres such as documentary, lecture-instructive,

historical, ephemeral, and ephemeral, with individual video durations ranging from 1 to 4 min and an overall duration of approximately 75 min. Similarly, the YT dataset consists of 50 MPEG-1 videos, also in colored with sound, with a overall size of 468 MB (Table 1).

Table 1. VS datasets details

Datasets	#Annotations	Categories	Duration (Min)	Frame Rate
Open Video [6]	5	documentary, educational ephemeral, historical, lecture	01–04	30 fps
Youtube [6]	5	Cartoons, news, Home, and	01–04	30 fps

5 Experimental Results and Discussion

This section evaluates the KVS model performance analysis as discussed in Sect. 4.

Quantitative Analysis. Quantitative analysis of VS needs to have standardized methods. We compare the performance of the KVS algorithm, which extracts keyframes from the videos, to user-generated summaries in the OV and YT datasets. We evaluate the effectiveness of these methods using the F-measure metric. On both datasets, we compare the results of our KVS approach to various keyframe selection-based VS approaches. Based on the analyses presented in Table 2 and Table 3, it is clear that our KVS approach achieves a better F-measure of 73.1% and 76.3% for the Open Video dataset and YouTube dataset, respectively, compared to other current approaches.

Table 2. Proposed KVS quantitative analysis on Open Video dataset

Algorithm	Precision	Recall	F-Measure
VSUMM [6]	48.2	63.1	55.0
SBVS [7]	74.14	70.3	72.1
EVS [12]	70.9	59.6	64.8
SUM-GANdpp [10]	–	–	72.0
ESVS [14]	69.4	61.8	65.4
DPCA + HSV [13]	66.6	60.6	63.4
Muhammad et al. [11]	–	–	67.0
Proposed KVS	**75.2**	**71.7**	**73.1**

Qualitative Analysis. The OV and YT datasets include ground-truth summaries created by five users for each of their 50 videos. We compared our proposed KVS methods with these user-generated summaries. Figure 2 presents the results of our experiment, showing a comparison between a sample video and the existing techniques.

Table 3. Proposed KVS quantitative analysis on YouTube dataset

Algorithm	Precision	Recall	F-Measure
VSUMM2 [6]	44.0	54.0	48.5
Jin et al. [15]	42.0	74.0	50.2
EVS [12]	53.0	49.7	51.3
DPCA + HSV [13]	74.4	64.0	68.8
VSUMM1 [6]	38.0	72.0	49.7
ESVS [14]	58.5	50.0	53.9
SUM-GANdpp [10]	–	–	60.1
Proposed KVS	**80.2**	**72.7**	**76.3**

Fig. 2. Qualitative analysis of KVS on Open video dataset

6 Conclusion

To produce static VS, this research introduces an efficient approach for a keyframe-based VS model. Our keyframe selection algorithm employs k-means clustering to identify keyframes based on the information they encapsulate. We extract keyframes using an unsupervised deep learning method that leverages k-means clustering. This technique uses feature fusion to extract features from ResNet-50 and GoogLeNet. After the fusion process, we apply k-means clustering to group frames, selecting the most representative frames from each cluster based on centroid calculations. Experimental results demonstrate that our proposed model performs effectively on two benchmark datasets. Exploring different combinations of additional pre-trained Dual-CNN models can further enhance the results. Future work will focus on determining the optimal number and combinations of pre-trained Dual-CNN models to improve the overall quality of video summaries.

References

1. Muhammad, K., et al.: DeepReS: a deep learning- based video summarization strategy for resource- constrained industrial surveillance scenarios. IEEE Trans. Ind. Inf. **16**(9), 5938–5947 (2019)
2. Khurana, K., Deshpande, U.: Two stream multi-layer convolutional network for keyframe-based video summarization. Multimedia Tools Appl., 1-42 (2023)
3. Dhiman, A., Deshmukh, M.: Optimized approach for video summarization using transfer learning and LSTM. In: 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, pp. 26-31 (2023). https://doi.org/10.1109/CISES58720.2023.10183585.
4. Negi, A., Kumar, K., Chauhan, P., Saini, P., Kashid, S.: Resource utilization tracking for fine-tuning based event detection and summarization over cloud. In: International Conference on Deep Learning, Artificial Intelligence and Robotics, pp. 73-83. Springer International Publishing, Cham (2022)
5. Negi, A., Kumar, K., Saini, P., Kashid, S.: Object detection based approach for an efficient video summarization with system statistics over cloud. In 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp. 1-6. IEEE (2022)
6. De Avila, S.E.F., Lopes, A.P.B., da Luz Jr, A., de Albuquerque Araújo, A.: VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognit. Lett. **32**(1), 56–68 (2011)
7. Kashid, S., Awasthi, L. K., Kumar, K., Saini, P.: NS4: a novel security approach for extracted video keyframes using secret sharing scheme. In: 2023 International Conference on Computer, Electronics and Electrical Engineering and their Applications (IC2E3), Srinagar Garhwal, India, pp. 1–6 (2023). https://doi.org/10.1109/IC2E357697.2023.10262778.
8. Saini, P., Berwal, K.: ESKVS: efficient and secure approach for keyframes-based video summarization framework. Multimedia Tools Appl., 1-29 (2024)
9. Saini, P., Kumar, K., Kashid, S., Saini, A., Negi, A.: Video summarization using deep learning techniques: a detailed analysis and investigation. Artif. Intell. Rev. **56**(11), 12347–12385 (2023)

10. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial LSTM networks. In: Proceedings of IEEE Conference Computing Vision Pattern Recognition (CVPR), Honolulu, HI, USA, vol. 1, pp. 2982–2991 (2017)
11. Asim et al.: A key frame based video summarization using color features. In: 2018 Colour and Visual Computing Symposium (CVCS), pp. 1–6. IEEE (2018)
12. Negi, A., Kumar, K., Saini, P., Kashid, S.: Object detection based approach for an efficient video summarization with system statistics over cloud. In: 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp. 1-6. IEEE (2022)
13. Asim et al.: A key frame based video summarization using color features. In: 2018 Colour and Visual Computing Symposium (CVCS), pp. 1-6. IEEE (2018)
14. Kumar, K., Shrimankar, D.D., Singh, N.: Equal partition based clustering approach for event summarization in videos. In: 2016 12th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), pp. 119-126. IEEE (2016)
15. Jin, H., Yang, Yu., Li, Y., Xiao, Z.: Network video summarization based on key frame extraction via super- pixel segmentation. Trans. Emerg. Telecommun. Technol. **33**(6), e3940 (2022)
16. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
17. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR (2015)
18. Nair, M.S., Mohan, J.: Static video summarization using multi-CNN with sparse autoencoder and random forest classifier. SIViP **15**, 735–742 (2021)
19. Otani, M., Nakashima, Y., Rahtu, E., Heikkila, J.: Rethinking the evaluation of video summaries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7596–7604 (2019)
20. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, pp. 445-450 (2016)