Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 26 (2023), No. 3, pp. 913–925
DOI : 10.47974/JDMSC-1778

Enhanced security using video summarization for surveillance system using deep LSTM model with K-means clustering technique

Shikha Sharma *
Department of CSE
Poornima University, Jaipur
Rajasthan
India

Dinesh Goyal †
Department of CSE
Poornima University, Jaipur
Rajastshan
India

Abstract

For the security of the Society, Video surveillance systems have become an important tool that can help in preventing crime in public or private establishments, leading to more safe and secure environment. With the rise of technology that can handle high volune video data and the need for continuous monitoring for security to citizens, the analysis of the video footages, during security breach attempts has become more challenging. Video summarization can play a crucial role in secure surveillance systems by optimizing the analysis time of the security breach. This research attempts to summarise the important aspects of the original video without losing its context and also optimizes the storage requirement. To achieve the same a unique KM-LSTM based technique has been developed that utilizes deep learning methods to understand links among highlight as well as non-highlight video segments through the use of pairwise distance calculations. A new Deep LSTM model has been used to extract the features from all segments of frames. These extracted features are clustered using k-means clustering. Then, such features are trained using a pairwise deep learning model to get highlight scores as well as leverage the comparative similarity among pairs of the highlight as well as non-highlight segments. At last, video summarization takes place to generate the video summary from the highlight segments only. From the results of these experiments discovering the moments of the user's major or special interest (namely, highlights) in a video, to generating summarization of videos. The above technique can help

* *E-mail:* shikha.sharma@poornima.edu.in (Corresponding Author)
† *E-mail:* dinesh.goyal@poornima.org

in increasing the analysis of Video content and thus achieving a more vigilant and secure society.

Subject Classification: *68M20, 68T07.*

Keywords: *Security, Video summarization, Deep learning, LSTM, TVSum50, K-means clustering, Pairwise distance calculation, Surveillance systems.*

1. Introduction

Video summarization is the process of condensing long video sequences into shorter, more manageable summaries while preserving the important details and events in the footage[1]. There are ample reasons why video summarization plays an important role to improve security- it is time efficient, it does bandwidth and storage optimization, easy information retrieval, Accessibility & Inclusion, Content Skimming, and Decision-Making & Analysis.

Video summarization helps security personnel to quickly identify and respond to security threats by condensing long video sequences into shorter, more manageable summaries while preserving important details and events. Video summarization improves situational awareness for security personnel, allowing them to make quick and informed decisions during emergencies. It aids in post-incident analysis by helping investigators to quickly review important events leading up to an incident and identifying key information that might have been missed in the original footage.The importance of video summarization in surveillance systems cannot be overemphasized. Here are some of the reasons why:

Efficient Use of Resources: Video summarization helps in making the most efficient use of limited resources such as bandwidth, storage space, and computing power. By summarizing video footage, only the most important segments are stored, reducing storage costs, and allowing for longer retention times.

Quick Identification of Suspicious Activities: Video summarization makes it easier for security personnel to quickly identify suspicious activities and respond to security threats in real-time. With video summaries, security personnel can scan through hours of footage in minutes, without having to watch every frame of the video.

Improved Situational Awareness: By providing a condensed version of the video footage, video summarization improves situational awareness for security personnel. This enables them to make quick and informed

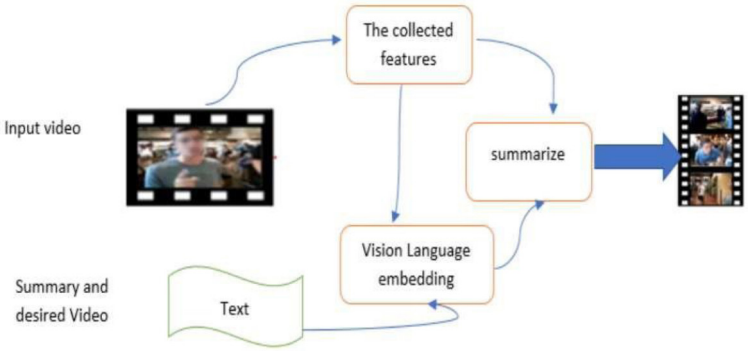

Figure 1

A video and optionally summarization procedures

decisions during emergency situations, leading to faster response times and improved safety.

Overall, Video summarization is an essential tool for ensuring the security of people and property in today's fast-paced world. It provides an efficient way of analyzing video footage, allowing security personnel to quickly identify and respond to security threats. With the help of video summarization, surveillance systems can be more effective and efficient in keeping people and property safe.[2] Figure 1 depicts a video summary method, which may or may not include the summation of the video.

Video summarization [3] is an instrument to create a short outline of a video to give a visual dynamic of the video arrangement to the user. For video summarization in recent years, Deep Learning (DL) has shown to be the most effective method available [4][5]. Deep learning approaches, on either hand, automatically analyze visual features from the data they are fed. And the use of machine learning (ML) is also becoming more popular among investigators in recent years. Due to the obvious cognitive structure of convolution neural network (CNN) deep learning model, it has been effective at achieving excellent facts in machine learning. [6].

Traditional DCNN-based approaches for video summarization are often associated with the following disadvantages. Using pretrained two-dimensional (2D) CNNs (VGGNet or GoogLeNet) to extract features, only carry spatial data of every frame, & as a result, the relationship between frames is not taken advantage. In this research, we present a unique approach (KM-LSTM) for video summarising to overcome shortcomings.

The following are the most significant contributions made by this research study:

- By measuring accuracy on the publicly available TVSum50 dataset, we established a new benchmark for video summarising. An end-to-end architecture is created by integrating feature extraction, keyframe production, and summary generation into a single technique.
- The suggested architecture obtains detailed and delicate video features via the use of an LSTM model to evaluate the value of frame-level feature importance. These characteristics are specifically learned for tasks to increase the accuracy of our technique.
- An improved loss function is created to constrain the derivative of sequential data & exploit possible sequential structure in video footage.
- To propose k-means clustering for keyframe generation and pair-wise calculation for highlights and non-highlights detection.
- To generate a platform for the summarization of any kind of video using the estimated frame-level feature importance.

The remaining parts are grouped in the following ways. Section 2 provides a review of the literature on the various options now available. Section 3 outlines the intended work as well as the steps necessary to complete it. Section 4 contains the results of the experimental examination. Section 5 gives the conclusion of implementation work also provides future scope.

2. Literature survey

Video summarization is a crucial tool for ensuring the safety and security of people and property in today's fast-paced world of surveillance.

Aboelenien et. al. [7] will examine video summarising approaches, with a particular emphasis on unsupervised approaches that rely on reinforcement learning or GANs, They also discovered that for certain approaches, while taking decent last F-score outcome. Hussain et. al. [8] has explained multiview video summarising (MVS) was accomplished by the integration of DNN-based soft computing methods in 2-tier architecture. Muhammad et. al. [9] designed deep CNN framework using hierarchical weighted fusion for summarizing of monitoring films. Xiong et. al. [10] developed a deep summarization network by an attention, this new video summarising framework dubbed attentive encoder-decoder networks for video summarization (AVS). Jabeen et. al. [11] suggested

a strategy for summarising multimedia data that is provided in the procedure of video. They first determine the borders of the scene by analyzing motion characteristics. And use CNN and bidirectional LSTM to eliminate redundancy of frames Ma M. et. al. [12] looked at the effect of video frame feature representation on the performance of dictionary selection-based VS. Deep features obtained using DNNs are initially utilized to characterize video sequence for dictionary selection-based VS. Therefore. Song S. et. al. [13] suggested a DRNN model that learns to extract category-driven keyframes from a set of input images. First, they used time-dependent location networking to successively extract a set number of critical frames from a video stream, then they used a RNN to categorize the video into one of many categories. Neha et. al. [14] shown importance of LSTM to generate summary of content. Gajanand et al [15] discussed the method involves comparing the difference matrices of the original and suspect videos, where the difference matrix is the absolute difference between two corresponding frames. Shikha et. al. [16] discussed importance of summarizing the online classes videos conducted during covid19. Girdhar et. al. [17] discussed various potential applications, including surveillance systems for public safety and security, healthcare monitoring, and activity recognition in sports and fitness.

3. Proposed methodology

This part focuses on the problem description and how to solve it using the technique offered. Our research strategy & flow diagram have been proposed.

A. *Problem Statement*

The importance of security in video generated by Closed-Circuit Television (CCTV) systems cannot be overstated. CCTV systems play a crucial role in enhancing security measures by monitoring and recording activities in various locations such as public spaces, businesses, homes, and institutions.

Huge data generated 24*7 from this CCTV. Video summarization techniques employ various methods such as keyframe extraction, object detection, action recognition, and semantic analysis to identify and capture the most salient and informative parts of a video. These techniques aid in condensing videos while ensuring that crucial information is preserved, ultimately enhancing the overall video viewing and analysis experience.

Table 1
Proposed LSTM Model

Model: "sequential_1"

Layer (type)	Output Shape	Param #
lstm_3 (LSTM)	(None, 32, 20)	4240
activation_3 (Activation)	(None, 32, 20)	0
dropout_2 (Dropout)	(None, 32, 20)	0
lstm_4 (LSTM)	(None, 32, 30)	6120
activation_4 (Activation)	(None, 32, 30)	0
dropout_3 (Dropout)	(None, 32, 30)	0
lstm_5 (LSTM)	(None, 40)	11360
activation_5 (Activation)	(None, 40)	0
flatten_1 (Flatten)	(None, 40)	0
dense_2 (Dense)	(None, 1024)	41984
dense_3 (Dense)	(None, 10)	10250

Total params: 73,954
Trainable params: 73,954
Non-trainable params: 0

B. *Proposed Methodology*

In this study, we investigate a generic video summarising technique based on keyframe extraction, which we call keyframe extraction. To construct summarised films, we offer a framework for extracting vision-based keyframes that may be used in conjunction with clustering or video skimming to create summarised videos. For our tests, we have utilized the TVsum50 dataset, which has been divided into several frames per second. LSTM model was used to extract features[18]. The extracted features from the frames are then used to train and later LSTM [19] is also used to classify the keyframes that are obtained by k-means clustering. Complete architecture of LSTM model is shown in Table 1. To obtain keyframes, we clustered the characteristics that were retrieved before. With this in mind, and taking into consideration that the enhanced score only wants to express a comparative degree of interest inside every video, we suggest training the extracted features from LSTM on each stream separately, which also characterizes the relative relationships between a set of pairs utilizing pairwise distance calculation. Each pair contains a highlights section from the very same video and a non-highlight piece from the same video.

Video skimming [20] is used to construct the highlight-driven analysis of a video by giving a highlight score to each section. Both of these methods maintain all of the parts in the movie while altering their speed

rates of playback following the highlight ratings. Only highlight portions are assembled, and non-highlight parts being deleted out of the sequence.

C. *Proposed Algorithm*

Requires TVSum50 data

1. Start
2. Collect the video data from the TVsum50 dataset.
3. Division of video data into several frames
4. Apply deep LSTM for feature extraction and select the best features from the frames
5. Feaure importance calculation
6. Divide the data between training, testing, as well as validation.
7. Train the model on extracted features from frames
8. Use k-means clustering for clustered the feature importance score of extracted features to find the similarity and dissimilarity and generate the keyframes
9. Utilizing pairwise distance computation, determine the degree of similarity among pairs of a highlight as well as non-highlight segments
10. Create a video summary by applying the video skimming method based on highlight scores
11. Generated video summary
12. Calculate the classification performance of the model using some evaluation matrices on test data
13. Obtain calculated results.
14. End

4. Experimental results and evaluation

The experiments that were carried out as part of this study are described in this section. Our deep features were analyzed and compared to a baseline technique to highlight the benefits of using our deep features in video summarising. It also gives video summaries that have been generated for each video, that we may use to compare our analyses with. Many parameters may be used to assess the outcomes of classification techniques, including the accuracy score, classification report, F1- score, confusion matrix, and so on.

A. *Dataset Description*

TVsum50 dataset: The TVSum50 dataset contains 50 movies obtained from YouTube, each of which has been annotated by a total of 20 individuals. It consists of 50 films of different genres (for example, news, how-to, documentary, vlog, egotistical) as well as 1,000 annotations with shot-level relevance ratings gathered via crowdsourcing efforts (20 per video). The video plus annotation data allows for an automated assessment of several video summary approaches, without the need to perform an (expensive) user survey on a large number of participants.

B. *Data Visualization*

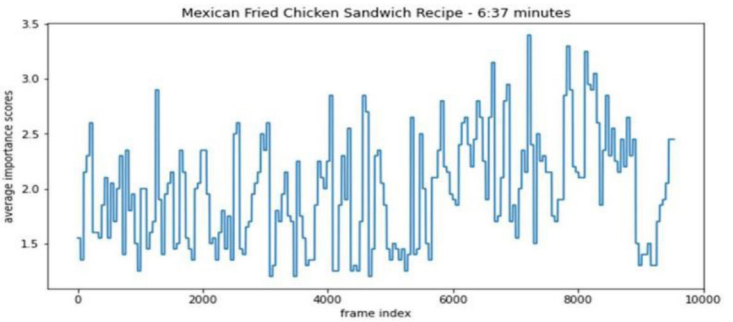

Figure 2

Data visualization graph of Frames Importance score graph

Above Figure 2 represent the data distribution of the Frames feature Importance score graph. This shows the frequency of Mexican fried chicken sandwich recipes within 6.37 minutes.

Figure 3

Highlighted frames on 10 Clusters

Figure 4

Highlighted Frames on 5 clusters

In above Figure 3 and 4 shows the highlighted frames on 10 and 5 clusters, respectively.

C. *Performance Evaluation*

To provide an accurate evaluation, we compare our suggested technique in a variety of contexts. For classification models, performance metrics are essential to computing the analytical results Figure 5 and 6 shown confusion matrix of proposed model.

Table 3 shown score of different performance parameters. Higher recall values indicating that the method was capable of extracting more of the needed frames, greater precision values imply fewer redundancies in the collected frames for the overview, whereas lower precision values suggest a larger number of duplicates. The last point to note about methods with a high F1-Score is that they are likely to have good recall as well as accuracy, which means they will have excellent overall effectiveness. Table 2 shown the performance of proposed model.

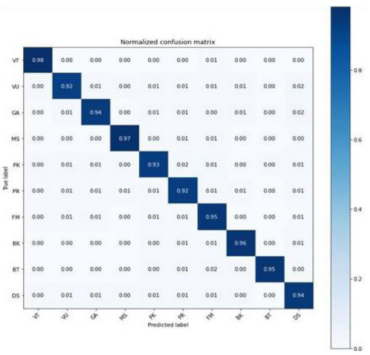

Figure 5

Normalized confusion matrix of the proposed model

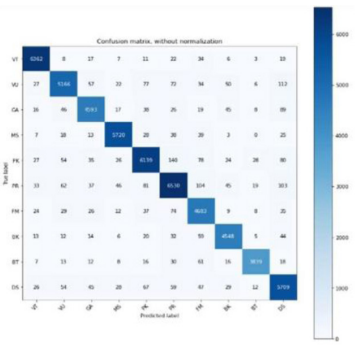

Figure 6

Without normalization confusion matrix of the proposed model.

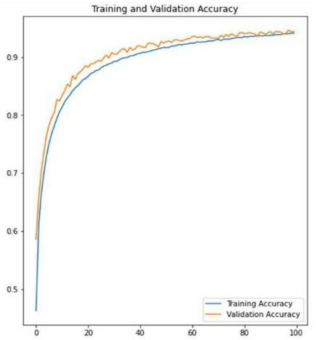

Figure 7

Traning and validation accuracy of LSTM model

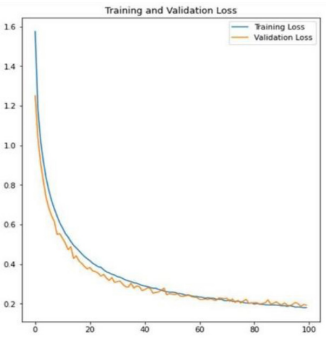

Figure 8

Training and Validation Loss over 100 epochs of LSTM model

Table 2

Performance of Training and validation of a proposed model

Model	Training loss	Training Acc	Validation Loss	Validation Acc
LSTM	0.1805	0.9410	0.1929	0.9439

Table 3

Parameters Performance of the proposed model

Performance Parameters		
F1 score	**Precision**	**Recall**
94.52	94.55	94.52

Table 4

Comparison results of Proposed and Baseline method

Method	Precision
Two Stream DCNN (baseline)	0.439
LSTM model (proposed)	0.9455

Figure 7 shown accuracy graph and figure 8 shown loss graph of proposed model. Table 4 shows the proposed model performance. From the experimented outcomes, it has been experimentally proven that LSTM perfect results as compared to the existing baseline method.

The accuracy of the technique is applied to estimate the total efficiency of the strategy. Whenever the TVSum50 dataset was utilized as the input, the LSTM Model method was only capable of delivering the highest level of accuracy (94.55 percent), according to the results. The baseline model achieved a score of just 0.439, which is much lower than our proposed model. The proposed model achieved higher performance in comparison to a baseline model.

5. Conclusion and future work

Video summarization is a critical aspect of modern-day surveillance systems that cannot be overlooked. By condensing hours of footage into manageable summaries, video summarization improves situational awareness for security personnel, allows for quick identification of suspicious activities, makes efficient use of resources, and aids in post-incident analysis. As a result, surveillance systems can be more effective in keeping people and property safe.

In this research, a video summarising strategy based on K-means and deep learning is created for automated video summarization. Our research investigates LSTM to build a unique supervised learning strategy to automate feature extraction from the frames of the TVSum50 dataset. In addition, to produce keyframes from the retrieved data, an unsupervised learning method called k-means clustering is employed. Our LSTM- based models beat competing approaches on difficult classification tests. This pairwise distance calculation is used to detect highlights from the video frames. In the last, video skimming is used to generate the video summary from the highlights keyframes.

The experiments results has shown the generated highlights for k=10 and k=5 clusters and also classification results of the deep LSTM model has been achieved good accuracy is 0.94 and minimal loss is 0.18 for the training set while the testing set is measured in terms of precision (94.55%), recall and f1-score are (94.52%). Also, the comparison has been done to validate the proposed work by comparing the baseline method (Two Stream DCNN) on precision parameters and found that the proposed deep LSTM based method achieved higher precision (94.55%) than it.

References

[1] Wesch, M. Digital ethnography: YouTube statistics (2008). On http:// mediatedcultures. net/ksudigg.

[2] Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., & Yokoya, N. Video summarization using deep semantic features. In *Asian Conference on Computer Vision*, pp. 361-377 (2016, November). Springer, Cham.

[3] Balas, D. M. & Agravat, S. J., Video Summarization using CNN and Clustering Algorithm. *Journal of Applied Science and Computations,* pp. 1249-1254 (2018).

[4] Haq, H. B. U., Asif, M., & Ahmad, M. B., Video summarization techniques: a review. *International Journal of Scientific & Technology Research*, 9, 146-153 (2020).

[5] Agyeman, R., Muhammad, R., & Choi, G. S., Soccer video summarization using deep learning. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 270-273 (2019, March). IEEE.

[6] Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q., Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6), e1264 (2018).

[7] Aboelenien, M., & Salem, M. A. M., Reinforcement Learning Versus Generative Adversarial Networks for Video Summarization. In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pp. 110-116 (2021, May). IEEE.

[8] Hussain, T., Muhammad, K., Ullah, A., Cao, Z., Baik, S. W., & de Albuquerque, V. H. C., Cloud-assisted multiview video summarization using CNN and bidirectional LSTM. *IEEE Transactions on Industrial Informatics*, 16(1), 77-86 (2019).

[9] Muhammad, K., Hussain, T., Tanveer, M., Sannino, G., & de Albuquerque, V. H. C., Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks. *IEEE Internet of Things Journal*, 7(5), 4455-4463 (2019).

[10] Ji, Z., Xiong, K., Pang, Y., & Li, X., Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6), 1709-1717 (2019).

[11] Khan, M. Z., Jabeen, S., ul Hassan, S., Hassan, M. A., & Khan, M. U. G., Video summarization using cnn and bidirectional lstm by utilizing scene boundary detection. In *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*, pp. 197-202 (2019, August). IEEE.

[12] Ma, M., Mei, S., Ji, J., Wan, S., Wang, Z., & Feng, D., Exploring the influence of feature representation for dictionary selection based

video summarization. In *2017 IEEE International Conference on Image Processing (ICIP),* pp. 2911-2915 (2017, September). IEEE.

[13] Song, X., Chen, K., Lei, J., Sun, L., Wang, Z., Xie, L., & Song, M. Category driven deep recurrent neural network for video summarization. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1-6 (2016, July). IEEE.

[14] Bansal, N., Sharma, A., & Singh, R. K. Recurrent neural network for abstractive summarization of documents. *Journal of Discrete Mathematical Sciences and Cryptography*, 23(1), 65-72 (2020).

[15] Sharma, G., & Goyal, D., Video tempering detection assessment in full reference mode using difference matrices. *Journal of Discrete Mathematical Sciences and Cryptography,* 22(4), 645-659 (2019).

[16] Sharma, S., & Saini, M. L. Analyzing the Need for Video Summarization for Online Classes Conducted During Covid-19 Lockdown. *In Data, Engineering and Applications: Select Proceedings of IDEA 2021*, pp. 333-342 (2022). Singapore: Springer Nature Singapore.

[17] Girdhar, P., Johri, P., & Virmani, D. Incept_LSTM: Accession for human activity concession in automatic surveillance. *Journal of Discrete Mathematical Sciences and Cryptography*, 25(8), 2259-2273 (2022).

[18] Deng, L., & Yu, D., Deep learning: methods and applications. *Foundations and trends® in signal processing*, 7(3-4), 197-387 (2014).

[19] Zhang, K., Chao, W. L., Sha, F., & Grauman, K., Video summarization with long short-term memory. In *European Conference on Computer Vision,* pp. 766-782, (2016, October). Springer, Cham.

[20] Zhang, L., Sun, L., Wang, W., & Tian, Y., KaaS: A standard framework proposal on video skimming. *IEEE Internet Computing*, 20(4), 54-59 (2016).