Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Condensation of Video for Activity Detection in CCTV Footage using Deep Learning

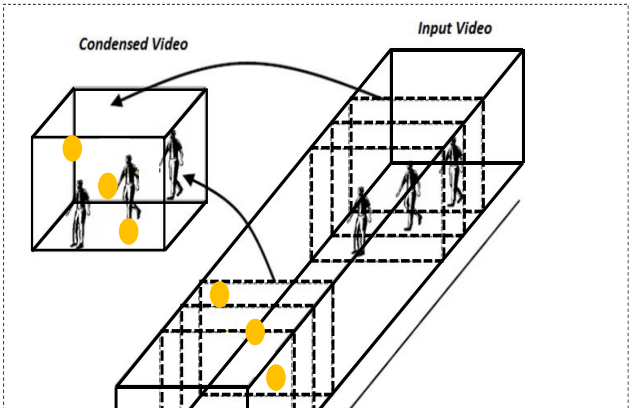
Suhandas,
A.J. Institute of Engineering & Technology Mangalore, India
Santhosh Kumar G
East West College of Engineering, Benagluru, India

Abstract—Video Surveillance is a very critical aspect from a security perspective for any organization or a residential compound. Video surveillance involves capturing video feed from multiple closed circuit television cameras. It is also preferable to have the video data in high resolution to better understand any incident of interest that is captured in the video footage. One of the major challenges with storing high resolution video footage is the volume of data available and the data storage space required. Another challenge posed to the users is the time spent in reviewing the video footage to identify the video segment of interest in order to identify an incident. Deep learning algorithms like Convolution Neural Network is an effective solution to recognize image data classification. Convolutional neural networks are utilized in this research to classify individual frames of video data to recognize if there is any activity in the frame or not. CCTV footage from a classroom was used to train and validate the proposed model. A validation accuracy of 99% was observed. The sequence of frames with activity are recognized and only those frames are combined together into a video file. The resulting video file is a condensed video which only shows the video segments with the activity that the model is trained to recognize.

***Keywords*—neural networks, machine learning, security, surveillance, video, condensation, deep learning, video tagging, classification**

I. Introduction

THE availability of low cost Closed Circuit Television CCTV systems, as well as high-speed internet, has resulted in a significant growth in the use of CCTV. The CCTV systems provide a sense of security to people when they go out of their residences and also to remotely monitor the premises and surrounding for any suspicious activities. CCTV systems continuously record video footages daily until the memory is filled up in the hard disk after which it starts deleting older data and recording new data. The amount of footage available will span over many hours. In case of any incident when the footage has to be reviewed, the user has to go through the entire footage to locate the incident in the footage. This is a very tedious process to browse and retrieve the video segment of interest. Efficient browsing and retrieval is a key tool in visual surveillance. Therefore, is a need to develop a methodology to aid human operators to identify relevant video segments and condense it to a video of interest. In this study a methodology is proposed to classify the video frames as frames with activity and frames with no activity. Convolution Neural Network has been gaining popularity with the availability of affordable high performance computing systems Convolutional Neural Network. Convolution Neural Network is a deep learning technique that is modeled after the human eye and can extract features from the images using layers of convolution and pooling. The retrieved features are sent into a fully connected network, which sorts the input characteristics into frame classes. The convolution and pooling layers extract features and provide a feature vector as an output. After that, the feature vector is flattened and fed into a fully connected dense layer, which performs the classification operation. The performance of deep learning network depends on the number of image samples available to train the model. Since the image data is extracted from a video footage, sufficient number of data samples are available to train the model. CNN is a deep learning model that can extract features



and classify them. A classroom's CCTV video was utilized to train and evaluate the model. The CNN model after training shall recognize the frames in which activity is present. In [1] the authors propose a deep summarization network to summarize a video. They offer an end-to-end reinforcement learning system with a new reward function that accounts for both variety and representativeness of produced summaries while avoiding the usage of labels or user input interaction. Deep summarization networks learn to create increasingly varied and representative summaries in order to maximize

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

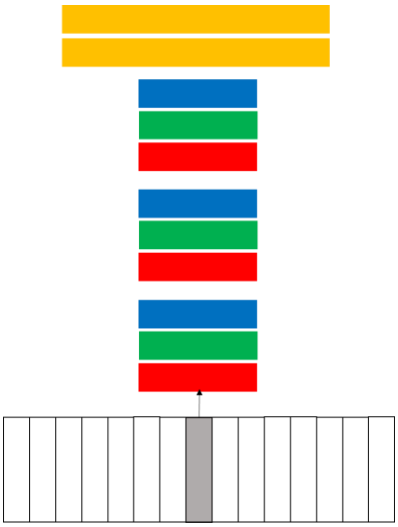


Fig. 2 Single Frame

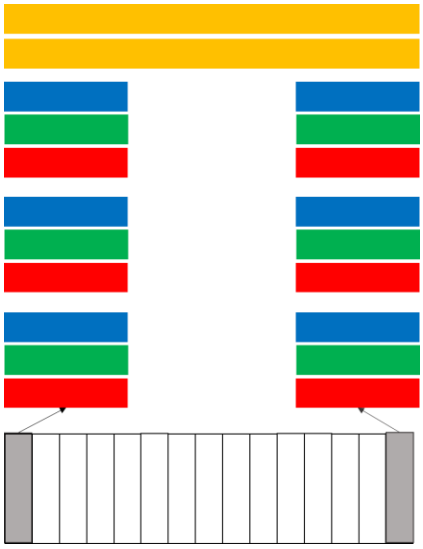


Fig. 3 Late Fusion

performing video summarization using unsupervised learning and standard vision-based algorithms for reliable feature extraction from video frames. To successfully summarize a film using interesting key frame extraction, they presented a deep learning-based feature extraction followed by several clustering algorithms. The deep learning-based feature extraction technique showed better performance. In [3], the authors propose a novel method that integrates the video scene ontology with convolution neural network for better video tagging. In this approach, the content of a video is captured by extracting information from key frames, after which the key frames are fed into a CNN-based deep learning model for training. The approach had a 99.8% total accuracy.

In [4], the researchers presented a new strategy for learning video summarization from unpaired data. The model learnt to create optimum video summaries utilising a collection of raw films and a set of summary videos with no relationship between the two sets, according to a deep learning framework

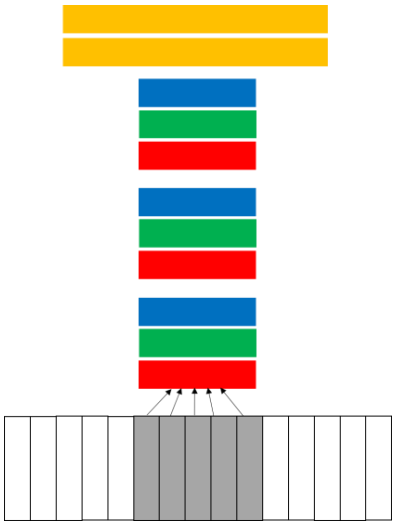


Fig. 4 Early Fusion

proposed. The authors developed a Detect-to-summarize

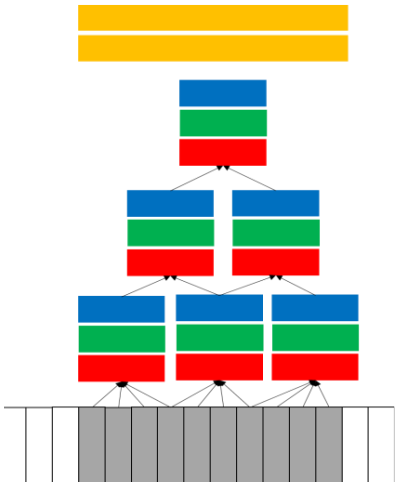


Fig. 5 Slow Fusion

Network architecture for supervised video summarising in [5], which featured both anchor-based and anchor-free methods. On the SumMe and TVSum datasets, the suggested technique outperformed most state-of-the-art supervised algorithms. In [6] proposes a deep learning system for detecting unforeseen accidents. A collection of event photos in tunnels was used to train a deep Learning model in the object Detection and Tracking System for Car, Person, and Fire, respectively. Accident films were used to evaluate the model. The authors suggested a deep learning-based strategy for detecting and classifying humans in video collected from distant distances using high-power lenses in [7]. A CNN Network with GoogLeNet Architecture was employed in this investigation. The method was able to detect persons with an accuracy of 90%.

To apply convolution neural networks for the classification of video data to account for temporal dependencies. In [8], four approaches were suggested. Single Frame model involves classifying videos by aggregation of class predictions of each

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

frame as shown in Fig.2. Late Fusion Model, combines frames by concatenating the first and last frame in the video segment as shown in Fig.3. Early Fusion model involves taking a larger segment from the video footage as shown in Fig.4. In Slow Fusion Model, partially overlapping segments are combined in successive Convolution Layers as shown in Fig.5. In this study, single frame model was considered and the model performed with good accuracy of 99.8%.

II. Objectives

The large volume of data acquired through 24x7 CCTV recordings of surveillance systems makes it cumbersome to store and view the recordings. In this study an activity-based video condensation approach is presented to achieve efficient browsing and retrieval of video segments of interest from the complete recordings. Single Frame Model is used in this study.

III. Methods

A. Data Acquisition and Pre-Processing

The footage captured by a classroom's CCTV system provided the data set needed to train, validate, and test the model described in this work. 6 Hours 30 Minutes of data was taken as the video data for training the CNN Model. Another footage of 8 hours 36 minutes was considered for validating and testing the CNN Model. As the proposed system is intended to be of use with CCTV footages, CCTV footage of one class room was chosen to train the model. Segments of Video where the classroom was closed and no activity was happening was identified and segments where there was activity in the classroom from when the classroom was opened till the class room was shut was noted. The video is given as input to the Frame Extractor module. The frames of the video

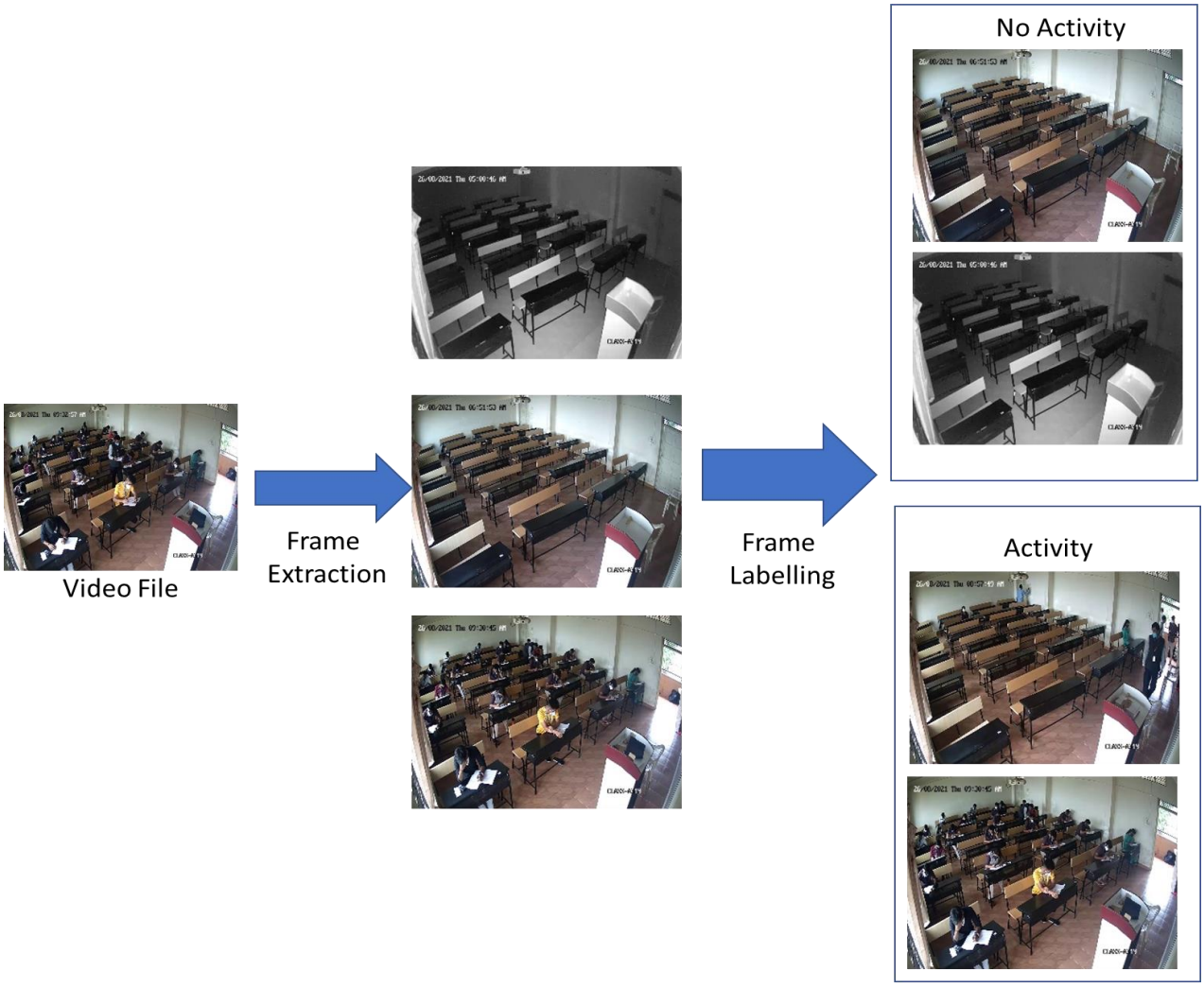


Fig. 6 Frame Extraction and Labelling

segments were extracted and the frames corresponding to the segments with no activity were labelled as 0 and the Frames corresponding to the segments with activity are labelled as 1. As shown in the fig. The data can now be used in the supervised training of the proposed deep learning model. As

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

(a)

(b)

Fig. 7 NO_ACTIVITY Data samples (a) Without Light (b) With Lights On

(a)

(b)

Fig. 8 ACTIVITY Data samples (a) Activity Starting (b) On-going-Activity

the model was trained with limited computing power, to deal with the challenge of running out of memory, one in every 10 frame was retained and the rest of the frames were discarded. The number of frames retained was 2335. . All of the frames are scaled to 224x224px and then normalised to keep the pixel values between 0 and 1. As illustrated in Fig. 6, the dataset is then divided into two parts: Training Dataset and Validation Dataset. 70% of the frames were utilised for training, while the remaining 30% were used to validate the proposed model. Fig. 7 (a) and Fig. 7(b) shows image samples labelled "NO_ACTIVITY". Fig. 8 (a) and Fig. 8(b) shows image samples labelled "ACTIVITY".

B. Feature Extraction and Classification

The proposed CNN model consists of a Feature Extraction Block which consists of 3 successive layers of Convolution Layer and a Max Pooling Layer as shown in the Fig. 9. 50% dropout is applied after every Max Pooling layer to mitigate over-fitting. The output of the feature extraction block is flattened and given as input to a Fully Connected Dense Layer which is a classifier block. 50% dropout is applied on the Fully

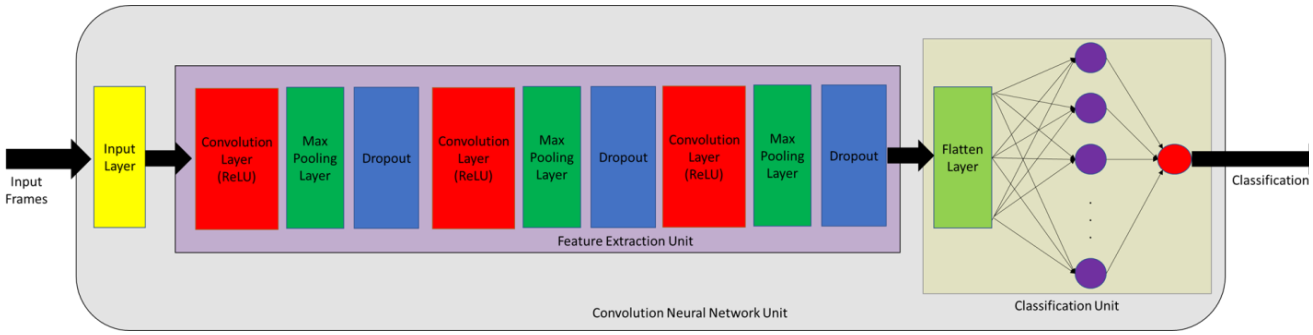


Fig. 9 Convolution Neural Network Model

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Connected network as well after the first dense layer. The final layer uses the Rectified Linear Unit (ReLU) Activation function, using softmax as its activation function, and is

TABLE I
CONVOLUTION NEURAL NETWORK MODEL

Layer (type)	Output Shape	Param#
Conv2d	(None,222,222,32)	896
MaxPooling2d	(None,111,111,32)	0
Dropout	(None,111,111,32)	0
Conv2d	(None, 109, 109, 64)	18496
MaxPooling2d	(None, 54, 54, 64)	0
Dropout	(None, 54, 54, 64)	0
Conv2d	(None, 52, 52, 64)	36928
MaxPooling2d	(None, 26, 26, 64)	0
Dropout	(None, 26, 26, 64)	
Flatten	(None, 43264)	0
Dense	(None, 64)	2768960
Dense	(None, 2)	130
Total Param	2825410	
Trainable Params	2825410	
Non-Trainable Params	0	

capable of categorising two classes, "NO ACTIVITY" labelled by 0 and "ACTIVITY" labelled by 1. Softmax function is suitable for multiclass classification. The Softmax function assigns decimal probabilities for each of the classes in the multiclass problem statement. The class with the highest probability is used to make the final classification. The suggested CNN Model is depicted in the diagram. TABLE I contains information on the layers. The model is trained for 10 epochs.

C. *Training and Model Evaluation*

The frames corresponding to the validation set are input to the model the frames are classified as 0 or 1. The classification of the frame is as shown is represented by [1,0] represents “NO_ACTIVITY” and [0,1] represents “ACTIVITY”. All frames corresponding to label 1 indicating activity is taken and condensed into a single video file for evaluation by the user.

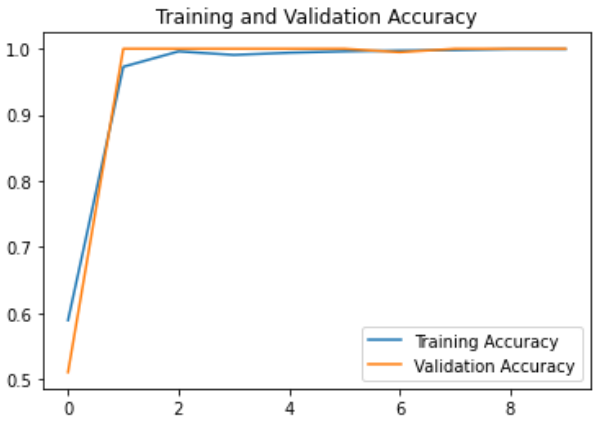


Fig. 10 Training and Validation Accuracy Curve

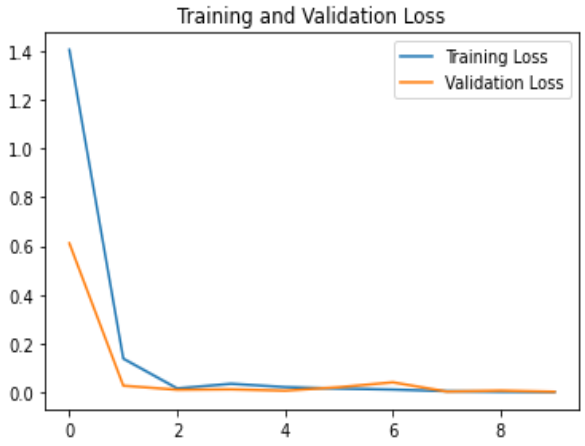


Fig. 11 Training and Validation Loss Curve

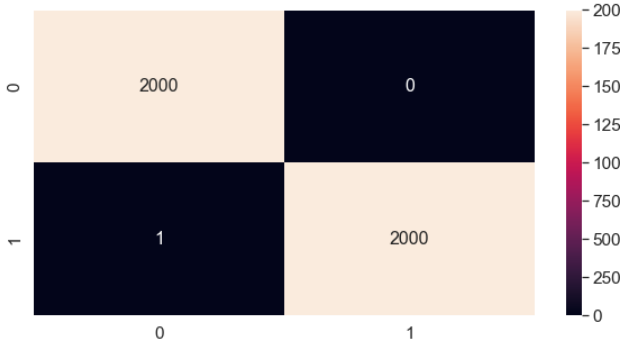


Fig. 12 Confusion Matrix for Validation Data

To reduce the computing load one in 10 frames was considered.

D. *Video Condensation*

The testing video is input to the system. Frame Extraction is done and the extracted frames are given as input to the CNN unit which classifies the frame as “ACTIVITY” or NO_ACTIVITY”. If the frame does not show any activity it is dropped. If the frame is classified as activity, then the frame is retained. All the retained frames which show activity are now aggregated in a frame aggregator module which will generate a condensed video file. The condensed video is a video file that only shows the segments of the footage involving activity.

IV. RESULTS

The model when evaluated with the validation set gave an accuracy of 99.8%. The accuracy curve for the model is as shown in Fig. 10. The Loss curve is as shown in Fig. 11. The fig 12 depicts the suggested model's confusion matrix. The proposed model accurately detects activity in video frames and can be used to identify only those frames with activity and condense the video into a video that only records the presence of activity, saving time and effort spent browsing through the entire video footage in search of some activity.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

V. Discussions

The proposed model shows promising results. The single frame technique was applied on the proposed model and the model showed a training accuracy of 99.9%. During validation and testing the model classified with an accuracy of 99.8%. Only one frame out of 4001 was misclassified in the validation confusion matrix, showing that the model has a high rate of True Positives (TP) and True Negatives (TN) and a low rate of False Positives (FP) and False Negatives (FN) (FN). Accuracy is calculated using equations (1).

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$

The accuracy for the validation data can be calculated from the confusion metrics. The accuracy is 99.9%

Conclusion

The proposed model with Single Frame Technique classified video frames with an accuracy of 99.9%. One in 10 frames were picked to reduce the load on the computing system. The Frame Extraction and training of the model was performed on a commercially available computing system with 32GB RAM, NVIDIA RTX 2060 GPU and an AMD Ryzen 7 Processor. This work can be extended to include CCTV footages from more classrooms and also to recognize specific type of activity or multiple activities.

Declaration

We have taken permission from competent authorities to use the data as given in the paper. In case of any dispute in the future, we shall be wholly responsible.

References

[1] K. Zhou, Y. Qiao and T. Xiang, “Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward” in arXiv, 2018, https://arxiv.org/abs/1801.00054

[2] S. Jadon and M. Jasim, “Unsupervised video summarization framework using keyframe extraction and video skimming”, in Proceedings of 2020 IEEE 5th International Conference on Computing Communication and Automation, pp. 140-145, 2020

[3] S. Ilyas and H.U. Rehman, “A Deep Learning based Approach for Precise Video Tagging” in Proceedings of 2019 15th International Conference on Emerging Technologies (ICET), pp. 1-6, 2019

[4] M. Rochan, Y. Wang, “Video Summarization by Learning From Unpaired Data” in Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 7894-7903, 2019

[5] W. Zhu, J. Lu, J. Li, J, Zhou, “DSNet: A Flexible Detect-to-Summarize Network for Video Summarization” in IEEE Transactions on Image Processing, Vol. 30, pp. 948-962, 2021

[6] K.B. Lee and H.S. Shin, “An application of a deep learning algorithm for automatic detection of unexpected accidents under bad CCTV monitoring conditions in tunnels”, in Proceedings of 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), pp. – 7-11, 2019

[7] H. Wei, M. Laszewski and N. Kehtarnavaz, “Deep Learning-Based Person Detection and Classification for Far Field Video Surveillance,” in Proceedings of 2018 IEEE 13th Dallas Circuits and Systems Conference (DCAS), pp. 1-4,2018

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks," in Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725-1732, 2014