International Journal of Advanced Technology and Engineering Exploration, Vol 9(95)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.875854

Research Article

An integrated framework for abnormal event detection and video summarization using deep learning

G. Balamurugan[1*] and J. Jayabharathy[2]
Research Scholar, Department of Computer Science and Engineering, Puducherry Technological University, Kalapet, Puducherry, India[1]
Associate Professor, Department of Computer Science and Engineering, Puducherry Technological University, Kalapet, Puducherry, India[2]

Received: 25-April-2022; Revised: 25-October-2022; Accepted: 27-October-2022

Abstract
In the real-world modern environment, intelligent transportation systems and intelligent surveillance systems are considered to play an anchor role in facilitating security and safety to the human society. These surveillance systems are diversely utilized in most of the places ranging from the application of border security to a street monitoring system that closely observes the abnormal event occurrence on the road. The core aim of the work is to present a rich set of abnormal event videos for intelligent surveillance systems. Hence, potential abnormal event detection and considerable video summarization mechanism is needed with maximum accuracy and reduced complexities. In this paper, hybrid convolution neural network (CNN) and bi-directional long short-term memory (Bi-LSTM) based abnormal event detection model is employed with reduced complexity. This model includes convolutional neural network with a pre-trained model for extracting spatio-temporal features from each individual frame selected from a series of frames, which is then passed to multi-layer Bi-LSTM that possesses the capability of accurately classifying the abnormal events in complex surveillance scenes of the road. Hierarchical temporal attention-based long short-term memory (LSTM) encoder-decoder model is included in the fine grain process in order to attain a better video summarization that preserves video key information and attains an optimal storage. The experimental results with the functional parameters confirmed the maximized abnormal event frames with minimum complexity using video summarization method and the accuracy of the abnormal detection scheme, on par with the benchmarked approaches considered for investigation.

Keywords
Road surveillance system, Abnormal event detection, Bi-directional long short-term memory, ResNet-50, Video. summarization, Hierarchical temporal attention-based long short-term memory encoder-decoder model.

1.Introduction
Video surveillance plays an emerging role for achieving intelligent transportation system for information extraction derived from the environment. The information extracted from real time scenario can be quite significant for detecting, classifying and summarization of abnormal events that includes investigations in video surveillance [1, 2]. A number of researchers over the decades concentrated on the objective of detection, classification and summarization of abnormal events in reduced computational complexities and maximized event frame [3, 4].

This field of research is considered to be essential for the enhancement of video surveillance systems that necessitates vast space of storage and analysis of complex data. Abnormal event detection, classification and video summarization is considered to be an important process that results in handling the challenges for improvement of video surveillance systems [5, 6].

The main issues in abnormal event detection and summarization is of immense concern to the research community in intelligent transportation systems [7, 8]. It is very difficult from the state-of-the-art techniques to provide the accurate detection and classification of abnormal events in the video frames and to summarize the key events. The core aim of the work is to develop the higher accuracy in detection,

*Author for correspondence

classification and summarization of abnormal events so as to enhance the efficiency for intelligent video surveillance. The detection of abnormal event concentrates on identifying the particular frames in a video that possesses anomaly. This abnormal event detection is determined to be the hot research issue in the video processing domain, since it provides maximized importance in surveillance system [9–12]. The enormous amount of data generated during monitoring process in surveillance system which makes the manual exploration of data associated with the frames completely infeasible. In this situation, the detection and classification of abnormal events present in the video frame sequence is also highly difficult [13–15]. Deep learning models are used for detection and classification process of abnormal event detection, since it plays an anchor role in the development of an automated abnormality detector that can handle the information associated with millions of video frames [16–19].Further, the process of video summarization becomes necessary for the principle of facilitating the information storage, simplifying the investigation of data and enhancing the access rate related to each individual video frame for obtaining maximized event frame content [20–23]. The video sequences permits the selection of key frames with abnormal events based on feature extraction and appropriate clustering schemes [24–27]. But, this selection of key frames can lead to redundancy resulting in meaningless frames that are essential to be removed through a required operation [28–31]. Thus, a deep encoder-decoder model is essential for video summarization as it plays a predominant role in preserving video key information and attaining an optimal storage. In this paper, hybrid convolutional neural network and bidirectional long short term memory are proposed for classifying events from each individual frame selected from the video sequences with minimum complexity. The proposed framework included the merits of hierarchical temporal attention-based long short-term memory (LSTM) encoder-decoder model for attaining superior video summarization that preserves video key information and attains an optimal storage. The experiments of the proposed framework are conducted using the datasets of performance evaluation of tracking and surveillance (PETS), unusual crowd activity dataset of university of Minnesota (UMN) and university of California San Diego (UCSD). In fact, to determine their predominance of the system it is compared to the benchmarked abnormal event detection and video summarization process with the performance metrics of accuracy, F1 score, precision and recall. The remaining sections of the paper are structured as follows section 2 presents the comprehensive review of the abnormal event detection model and video summarization schemes propounded in the literature over the recent years with the merits and limitations. Section 3 describes the proposed work of convolutional neural network and bi- directional long short term memory used in the process of abnormal event detection. Section 4 discussed the procedures required in applying the LSTM encoder-decoder model with hierarchical temporal attention to the process of significant video summarization. Section 5 demonstrates the experimental results and discussion of the proposed framework on par with baseline schemes with respect to the datasets of PETS, UMN and UCSD, respectively. The conclusion, major contributions and the future scope of this research is presented in section 6.

2.Literature review

Chu et al. [32] proposed spatio-temporal information for enhancing the degree of accuracy for input which is derived based on the inclusion of three-dimensional CNN that facilitates better feature extraction. It was proposed for handling the capability of training the CNN network, even without the utilization of category labels. This method applied the sparse coding over the manually crafted features derived from the inputs in order to support the unsupervised feature learning. It included a multi-level similarity associated between the inputs based on the shared statistical information. It also included the concept of quadruplet for modeling the structure of multi-level similarity, which was consequently utilized for a generalized triplet loss with the objective of training the CNN network [33–35]. It derived rich and robust feature representations by jointly optimizing the unsupervised feature learning and sparse coding. This detection scheme attained its objective based on the computation of sparse reconstruction error in order to predict the anomaly score associated with each individual input frame. The experimental results of the detection scheme confirmed better accuracy and reduced time in training, compared to the state-of-the-art methods [36–38]. Then, a feature expectation sub-graph calibrating classification method-based abnormal event detection scheme was proposed by Ye et al. [39] for attaining better accuracy in classifying the consecutive video frames. This method constructed feature expectation sub-graphs for the purpose of calibrating the classification purposed of the utilized classifier. It employed CNN and long short-term memory models for extracting the spatio-temporal

features of the video frames. It also concentrated on building the feature expectation sub-graph for every individual key frame of each video through CNN, such that it could be utilized for capturing topological and internal sequential associations of the structured feature vector[40]. It projected the sparse vector over the expectation sub-graphs for integrating them with the support vector machine for the objective of calibrating linear support vector classifier. The accuracy of this detection was determined to be maximized and the training time was minimized, compared to the baseline approaches.

Further, an abnormal event detection scheme using bidirectional multi-scale aggregation networks (BMAN) was proposed by Lee et al. [41] for obtaining improved accuracy with reduced training time. This detection approach included BMAN for learning spatio-temporal patterns for detecting changes associated with the learned normal patterns in order to confirm abnormalities in video frame sequences [42–44]. It included the merits of an appearance-motion detector and inter-frame predictor for identifying the deviation between consecutive video frames. An attention-based BMAN is included for generating an inter-frame that aids in encoding normal patterns through inter-frame predictor. It attained normal pattern encoding for complex motions and object scale variations based on the process of feature aggregation. The events of abnormality are detected by appearance motion detection based on encoded normal patterns that determined both the motion and appearance features for abnormality event identification. An adversarial three dimensional convolutional auto-encoder-based abnormal event detection approach was proposed by Song et al. [45] for achieving maximized accuracy by learning motion and appearance features simultaneously. It was proposed for determining the fine-grained spatio-temporal patterns. It was developed for learning the spatio-temporal features that aided in identifying abnormal events that distinguishes them from learning normal patterns from video frames. The encoder included in this approach captures low-level associations between the temporal and spatial dimensions of video frames. It also generated unique features that aided in representing information that portrays spatio-temporal information from the consecutive video. In addition, Yan et al. [46] propounded an abnormal event detection approach using semi-supervised learning for maximized accuracy in detection. This semi-supervised scheme composed the data that inherited the model that handled the two-stream structure. This two stream structure possessed motion streams and appearance with information, which is not degraded at any level of exploration. It included a recurrent variation auto-encoder that handled each individual stream in order to model them based on the merits of the normal distribution. The motion properties derived by this model aided in facilitating complimentary information that expresses the probabilistic distribution. The experimental validation of this model conducted using baseline datasets confirmed better performance on par with the benchmarked approaches.

A video summarization method-based on local alignment was proposed by Huang and Wang [47] for summarizing events associated with the multi-view videos in order to derive an optimal solution. It included deep learning strategy for feature extraction that handled the deviation problem of illumination and eliminate the details of fine texture. It also detected objects in the frame with the maximized capability. It incorporated features from accelerated segment test algorithm for enabling interview dependencies possible between videos' multiple views based on the concept of local alignment. It derived the benefits of object tracking for lower activity frames extraction. This video summarization method potentially minimized video content and, at the same time kept the information momentum in the events pattern. Another, video summarization with the steps of intelligent video capturing, removal of coarse and fine redundancy was proposed by Dilawari and Khan [48] for the better summarization process. It included resource-restricted devices for video capturing the events that pertain to industrial internet of things network. It employed the process of removing coarse redundancy based on the comparison of low-level features. It also facilitated the selection of candidate key frames and refined key frames for the process of differentiating them as the part of the summary.

The various methods are employed for the analysis of abnormal events to provide accurate accident content from the stack of videos. It is observed that tracking the vehicle in surveillance videos and video summarization techniques for abnormal event detection are considered as challenging part for generation of accurate accident events from the stack of surveillance videos. The core aim of the proposed work is to progress higher accuracy in detection, classification and summarization of abnormal events so as to enhance the efficiency for intelligent video surveillance.

3.Proposed method

In this paper, an abnormal event detection approach based on hybrid CNN and LSTM was proposed for extracting motion and learning normal appearances simultaneously in order to capture potential normal spatial and temporal patterns. The key elements of this abnormal event detection approach are presented in *Figure 1*. This abnormal event detection process comprises of three main steps such as, i) extracting features using pre-trained ResNet-50 model from the surveillance video, ii) generating feature vector from the consecutive fifteen frames derived from the video and, iii) multi-layer CNN-LSTM is used for recognizing anomalous events from the determined feature vector.

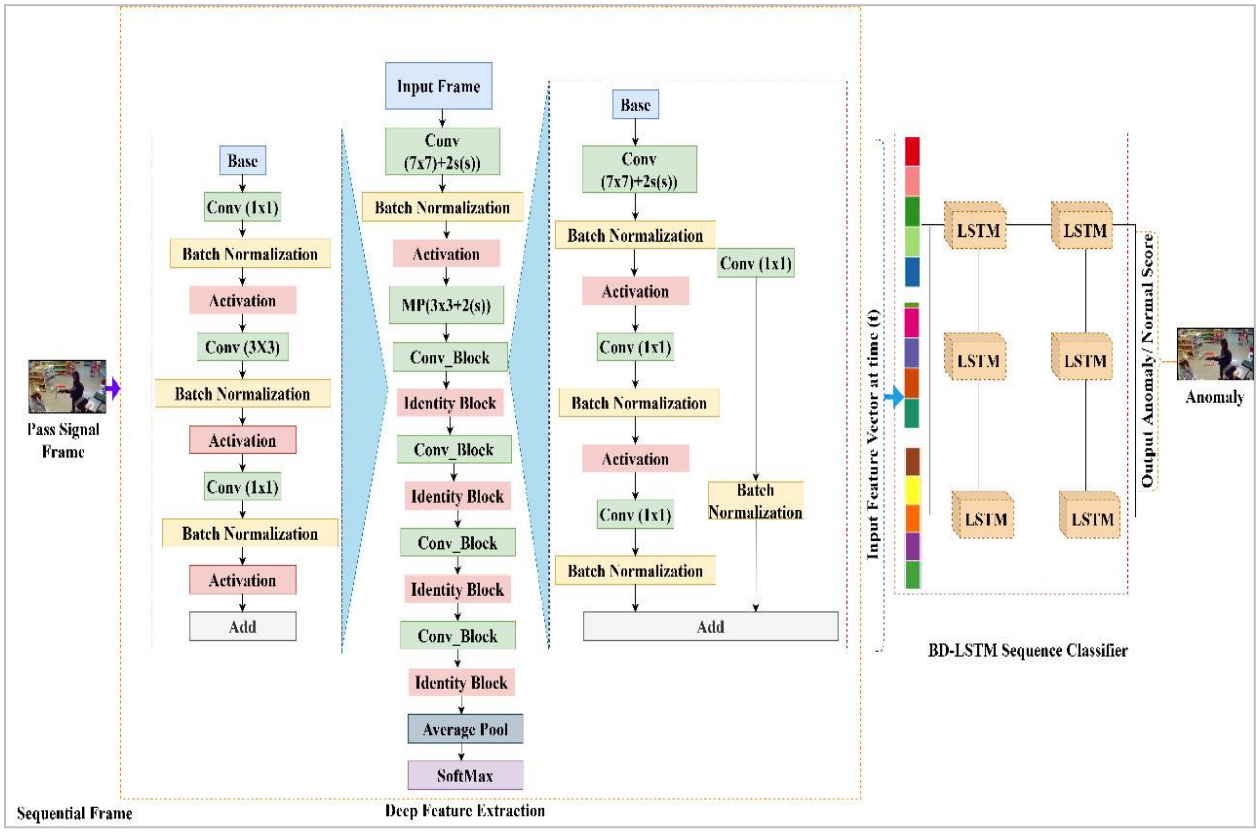

Figure 1 Abnormal event detection using CNN and bi-directional long short-term memory (Bi-LSTM) deep learning model

3.1Extracting features using pre-trained ResNet-50 model from the surveillance video

In the primary step, the pre-trained ResNet-50 is utilized, since it is a deep learning model that need a large quantity of images for training them from scratch. This ResNet-50 possesses high capacity processing units as they are essential for achieving better training at the primary level. In this abnormal detection framework, the ResNet-50 residual networks that are pre-trained are utilized for feature extraction.

This process of training ResNet-50 residual networks is completely based on the UMN dataset. In specific, ResNet-50 refers to one of the potential deep CNN models that utilized shortcut connections in which one or more layers are conditionally skipped in a predominant manner. In this ResNet-50 CNN model, the primitive cell blocks are defined as bottlenecks and complies based on the following rules such as, i) the amount of filters are doubled up when the map feature size is reduced and ii) equal likely number of filters in each layer is considered to possess an corresponding number of output feature maps. Then, the process of down-sampling is applied based on two strode convolutional layers, which is subsequently followed by the batch normalization prior. This down-sampling process is employed before the application of the Rectified Linear Unit (ReLU) activation function. Then, the method of

shortcut projection is employed, when the input and output dimensions are different. This projection method is mainly employed in fitting up the dimensions with the 1 × 1 convolutions. When the dimensions are same, then the identity shortcut is employed. The employed ResNet-50 possesses a 50 number of weighted layers with maximum of 23.5 million trainable parameters. The features are extracted and derived from the last fully connected layers of ResNet-50 comprises of 15-frames sequence. This sequence is further fed into the LSTM for processing. The ResNet-50 structural design is depicted in *Figure 2*.

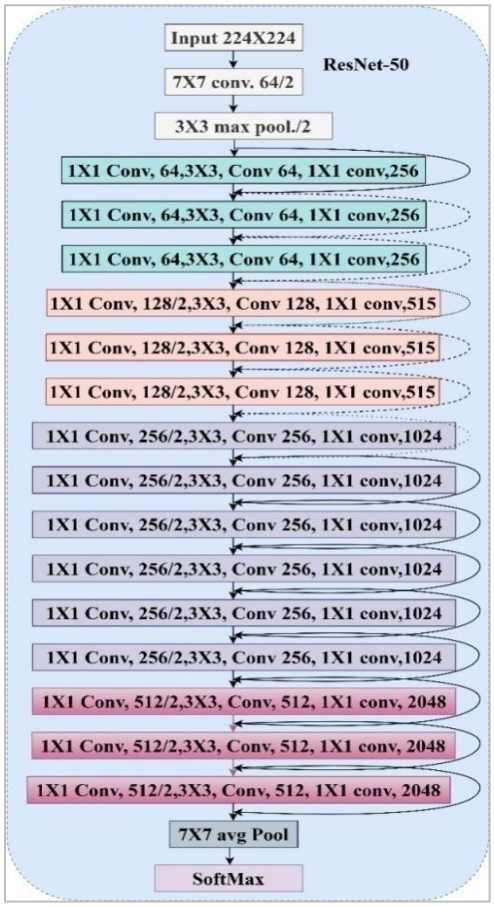

Figure 2 ResNet-50 structural design used in the abnormal event detection

The output of the CNN is given as input to Bi-LSTM as they are capable in exploring the concealed sequential information present in temporal and spatial temporal data. A video is considered to possess a sequence of frames that facilitates potential information for recognizing the context sensitivity of the event. In this context, recurrent neural networks are determined to be capable in the significant reading of the video sequence. However, the performance of recurrent neural network crumbles when the video sequence is comparatively long. This is mainly due to the reason that they fail to remember the earlier patterns of the video sequence. This limitation of recurrent neural network (RNN) is termed as the problem of vanishing gradient. This limitation of recurrent neural network can be handled through the use of LSTM, since it is capable of tracking longer sequences with maximized accuracy and optimality. This LSTM possesses a block with the gates of input, forget and output, which is significant in regulating pattern recognition of long-term video sequences. It specifically utilizes the Sigmoid function units as the core gates' component, which is acquired over the entire over the training duration. The operations and data flow of the video sequences handled by the LSTM units from the input to the output gates are presented in Equation (1 to 7).

$$I_{g(t)} = \sigma(\alpha_i[Y^t + R_{t-1}] + \beta_i) \quad (1)$$

$$F_{g(t)} = \sigma(\alpha_F[Y^t + R_{t-1}] + \beta_F) \quad (2)$$

$$O_{g(t)} = \sigma(\alpha_o[Y^t + R_{t-1}] + \beta_o) \quad (3)$$

$$S = tan(\alpha_S[Y^t + R_{t-1}])) + \beta_S \quad (4)$$

$$B_t = B_{t-1} * F_{g(t)} + S.* I_{g(t)} \quad (5)$$

$$R_t = tan(B_t) * O_{g(t)} \quad (6)$$

$$S_{Pred} = Soft\_Max(V_{R(t)}) \quad (7)$$

Where, 'i' refers to the input determined over time 't'. In this case, the Sigmoid function is depicted through $\sigma$with terms α and β representing the weights associated with the stage of training, respectively. Moreover, the terms $I_{g(t)}$, $F_{g(t)}$and $O_{g(t)}$highlights the input, forget and output gate of the LSTM at any time 't'. The input gate $I_{g(t)}$is responsible for monitoring and determines the time at which recording of the current input can be attained at the time 't'. The output gate $O_{g(t)}$verifies whether the data is shifted from the current memory B_t to its hidden state. The $F_{g(t)}$ forget gate estimates the time during which the preceding memory cell B_{t-1}can be released. However, the final detection of abnormal events is attained through the inclusion of SoftMax classifier, since it does not necessitate the transitional output associated with the LSTM. In contrast to the traditional LSTM, the bi-directional LSTM depends on the predecessor and successor video frames of the sequence. In addition, the bi-directional LSTM is determined to be relatively simple with the integration of double stacked recurrent neural networks. Among the double stacked RNNs, the first one is in the backward direction and second one is in the forward direction. Then, the hidden states related

to the double-stacked RNNs are integrated in the output. In this proposed abnormal event detection approach, a multiple layer bi-directional LSTM is used with each individual layer comprising of two cells that are responsible for achieving the forward and backward pass.

3.2Hierarchical temporal attention-based LSTM encoder-decoder model-based video summarization

This process of video summarization using hierarchical temporal attention-based LSTM encoder-decoder model comprises of four significant steps, such as, i) data acquisition based on vision sensor integrated with resource-limited device, ii) low-level features-based coarse refinement of derived frames determined from resource-limited device, iii) production of candidate key frames using the concept of sequence learning for generating final video summarization. In addition to the aforementioned three steps, a query from the user in order to generate the necessitated number of key frames that ultimately derives the compact video summary. *Figure 3* presents the complete process and steps involved in the implementation of the proposed hierarchical temporal attention-based LSTM encoder-decoder model-based video summarization approach is discussed as follows.

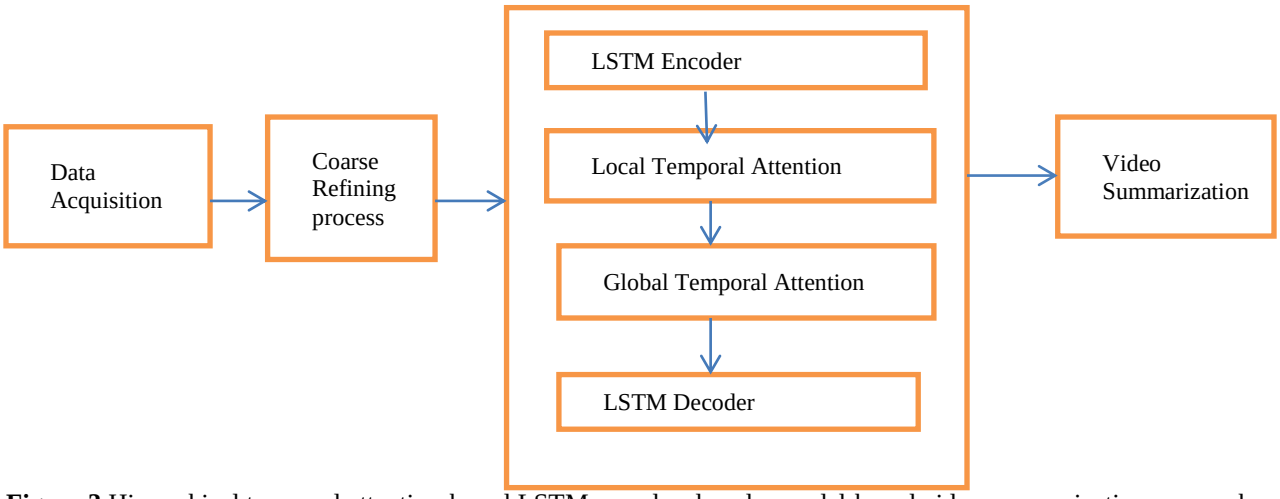

Figure 3 Hierarchical temporal attention-based LSTM encoder-decoder model-based video summarization approach

3.2.1Process of coarse refining

This coarse refining process considered the input frames from the vision sensor and derived as a collection of selected frames with comparatively reduced redundancy. The mechanism for feature comparison with reduced computation complexity is used in this coarse refining process. In this coarse refining process, oriented features from accelerated segment test and rotated oriented features are used as the superior representative for differentiating consecutive frames and low-level features. This possible utilization of oriented and rotated brief features are mainly due to its predominant performance compared to speeded up robust features, scale-invariant feature transform (SIFT) and the histogram oriented gradient (HOG) approaches. In this coarse refining process, the local features are extracted from two consecutive frames. Then, the feature vectors are compared in order to determine whether both the frames pertain to the same or dissimilar events of the video. At this juncture, the features frames belong to the same events, when the mutual information computed between the features is lesser than the output value. On the other hand, the features frame belongs to dissimilar events, when the mutual information calculated between the features is greater than the output value. This method of mutual information is used rather than the simple distance formula, since they aid in better understanding about the difference between events. In this proposed scheme, the optimal threshold is selected to be 1.94 (as determined as optimal after several numbers of experiments). Finally, each and every single frame representing each individual event is forwarded to the subsequent step of fine refinement.

3.2.2Process of fine refining

In this process of fine refining, the coarse refined frames are compared with the other frames based on sequence learning achieved through the hierarchical temporal attention networks. The primary goal of this fine refining process targets on deriving output candidate's key frames with higher accuracy. It is

visualized that in most of the methods, frame-level processing is considered for preventing redundancy by comparing the successive frames. However, extra computation is required in the frame-level process. Likewise, diversity is inadequate during the process of video summary generation, since the information associated with different events at the frame-level is considered to be missing. The aforementioned limitations are handled through the inclusion of hierarchical temporal attention networks.

3.2.3Use of hierarchical temporal attention networks

Hierarchical temporal attention networks used in the process of fine graining possess two layers that include, i) global temporal attention and ii) local temporal attention. The global temporal attention completely concentrates on the relative significance of different key frames with a candidate key frame. On the other hand, the local temporal attention approach calculated the relative potential of different frames within each key frame. The assignment of weights to t-length frames of a key frame is attained in two stages, rather than in a single stage in the single-layer temporal attention mechanism.

The global temporal attention determines the weighted summation vector of a key frame. In this case, is the frame-level vector that cumulates the complete set of correlated information for previous video key frames for determining the target key frame prediction based on the below Equation 8.

$$CI_{FL(Vector)} = \sum_{d=1}^{N_d} v_{tf(d)} X \; G_{tf(d)} \quad (8)$$

Where, $CI_{FL(vector)}$ is the correlated information vector that combines the complete amount of associated information from the key frames ($G_{tf(d)}$) and individual key frame density ($v_{tf(d)}$) in order to concentrate on the target location prediction process using equation 9.

$$v_{tf(d)} = W_d Tan\square(X_a^t[g_{t-v}; d_{t-v}] + X_a^t G_{tf(d)} + C_b) \quad (9)$$

$$v_{tf(d)} = \frac{e^{v_{tf(d)}}}{\sum_{d=1}^{N_d} v_{tf(d)}} \quad (10)$$

Where, $W_d$and C_b represents the weight associated with the key frame and value of bias considered in determining the correlated information. Further, $G_{tf(d)}$and X_a^t highlights the group training factor and individual vector considered for estimating correlation. In addition, is the number of vectors considered in evaluating the key frames of the complete video.

3.2.4Generation of video summary

This generation of summary step necessitates the user, whose individual inputs are constructed as a query which is related to the required number of key frames that are present within the threshold of candidate key frames. The final required key frames are chosen from the candidate key frames that portrays the sequence of occurred events. The number key frames selected from the candidate key frames completely depend on the user's requirement. This is because, they help in sustaining the candidate key frames that highlight the actual skims of the complete video. In specific, the method of multi-view summarization is employed for estimating the parameters of information computing in the case of key frame extraction necessitated for the users. Further, the entropy associated with candidate key frame is computed for arranging them in the ascending order. This step extracts the frames with a highest entropy score based on the user input requirement that concentrates on the number of key frames. In this case, the information existing inside each and every frame indicates its level of entropy. Moreover, the frames with maximum information are considered to possess highest entropy score within it.

4.Experimental results and discussion

The performance of the proposed abnormal event detection scheme and video summarization scheme is investigated and compared with the benchmarked schemes based on experiments conducted on a server system that consists Nvidia TITAN XP GPU, 128 GB memory and Intel Xeon Scalable Silver 4114 CPU @ 2.20 GHz. The implementation of the abnormal event detection scheme and video summarization scheme are achieved using TensorFlow. The performance of the proposed framework is evaluated based on the public datasets of UMN, PETS and UCSD. During the training process, the datasets are considered to contain only normal events. On the other hand, the validation set includes both normal and abnormal events.

4.1UMN datasets

UMN dataset comprises of three different scenes that includes 11 video clips. In the UMN dataset, the initial 200 frames associated with each individual clip are considered as the training set and the rest frames corresponds to the validation set [23]. Totally, the testing set and training set consists of 4, 439 and 3, 300 frames, respectively. Moreover, the pixels comprise of frame resolution 320x240.

4.2UCSD dataset

This UCSD dataset in turn consists of two sub-datasets such as Ped 1 and Ped2 for recording the walkways of pedestrians [24]. Generally, Ped2

dataset is considered for most of the evaluation, since I have higher resolution compared to its counterpart. Ped2 datasets are also considered to possess abnormal events of ambiguity and corruption of data frames. These two sub-datasets consist of abnormal events that consist of scenes associated with cars, carts, bicycling and skating among the pedestrians [25, 27]. The Ped2 dataset of USCD includes 16 video clips and 12 video clips for the purpose of training and testing with the frame resolution of 360×240 pixels. Hence, a total of 2, 010 and 2,550 frames are present in the testing and training set.

4.3PET dataset

PET's data set provides the assessment on-board surveillance systems for the surveillance in roads. It includes the Dataset like ARENA is used for surveillance in multi sensor application. The datasets allow for detection of various surveillance events ranging from abnormalities to dangerous and criminal situations. The characteristics of ARENA dataset are employed as follows with the model of BIP2-1300c-dn and given resolution: 1280×960 pixels; frame rate: 30 fps. PET dataset is concentrate on event detection in crowded scenarios.

The evaluation metrics considered in determining the performance of the proposed mode; are accuracy in classification and F1-score. Accuracy is defined as the ratio between true positive and true negative to the total (true positive, true negative, false positive and false negative) determined during the classification process as depicted in Equation 11. F1-Score is defined based on precision and recall (Equation 12 to14).

$$Accuracy = \frac{True_{Positive} + True\_Negative}{Total(True\_Positive + True\_Negative + False\_Positive + False\_Negative)} \quad (11)$$

$$F1 - Score = 2 \times \frac{Pr\,ecision \times Re\,call}{Pr\,ecision + Re\,call} \quad (12)$$

$$Pr\,e\,cision = \frac{True\_Positive}{True\_Positive + False\_Positive} \quad (13)$$

$$Re\,c\,all = \frac{True\_Positive}{True\_Positive + False\_Negative} \quad (14)$$

In this context, the equal error rate of abnormal event detection system is attained, when the operating threshold for the accept/reject decision is adjusted such that the probability of false acceptance and that of false rejection become equal. Further, the performance of the proposed CNN-LSTM model applied over the datasets of UMN, PETs and USCD datasets are evaluated based on two metrics such as frame-level criterion and pixel level criterion. Initially, the criterion of frame-level is utilized for testing how better the detection of abnormal frames is achieved. The evaluation conducted using frame-level criterion clearly proved that the score of abnormality pertaining to normal scenes are considerably very low. On the other hand, it is also proved that the score of abnormality pertaining to abnormal scenes are considerably very high. Then, a frame-level receiver operating characteristic curve (ROC) is determined based on the estimated abnormality score with respect to true positive and false positive rate by varying the thresholds within the range score of abnormality. In addition, the area under the curve value associated with the receiver operating characteristic curve is estimated and compared with the values related to the state-of-the art schemes used for comparison. The results of the area under curve at the frame-level attained by the proposed CNN-LSTM model and benchmarked abnormal event detection schemes of the literature with respect to the datasets of UMN, PETs and USCD are presented in *Table 1*, *2* and *3*, respectively.

Table 1 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect to frame-level Area Under the ROC Curve (AUC) based on UMN dataset

Techniques	**AUC at the frame level (in %)**
Sparse Coding	94.28
Subgraph Calibration	95.39
Multi-Scale Aggregation	97.18
CNN-BiLSTM	**98.18**

Table 2 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect to frame-level AUC based on PETS dataset

Techniques	**AUC at the frame level (in %)**
Sparse Coding	96.83
Subgraph Calibration	97.12
Multi-Scale Aggregation	97.39
CNN-BiLSTM	**98.24**

Table 3 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect to frame-level AUC based on UCSD dataset

Techniques	AUC at the frame level (in %)
Sparse Coding	96.24
Subgraph Calibration	96.78
Multi-Scale Aggregation	96.16
CNN-BiLSTM	**98.36**

Furthermore, the performance of the proposed CNN-LSTM model applied over the datasets of UMN, PETs and USCD datasets are evaluated based on pixel level criterion. This criterion of pixel-level is included for evaluating how better the abnormal regions are potentially localized in the scenes considered for exploration. This proposed model included the merits of appearance-motion joint difference map that possesses the local information that are inherent to the detected abnormal events. But, the pixels that are activated in the scene frame are determined to be scattered sporadically. Hence, the included different map is partitioned into overlapping patches. Then, the mean of each overlapping patch associated with each image plane is used for clustering the activated neighborhood pixel for the localized information related to the detected events. At this context, if the detected regions envelop more than 40% of the abnormal ground truth pixels, then the regions is considered to the true positive frame. Otherwise, they are identified as the false positive frame. The ROC curve at the pixel-level is determined by varying the threshold, such that the value of AUC determined from the pixel-level ROC curve can be utilized for performance comparison. The results of the AUC at the pixel-level attained by the proposed CNN-LSTM model and benchmarked abnormal event detection schemes of the literature with respect to the datasets of UMN, PETs and USCD are presented in *Table 4*, *5* and *6*, respectively.

In the second fold of investigation, the performance of the proposed CNN-LSTM model is evaluated based on detection time incurred per frame (seconds)with respect to the datasets of UMN, PETs and USCD are presented in *Table 7*, *8* and *9*, respectively. The detection time incurred per frame (seconds) by the proposed CNN-LSTM model is confirmed to be always lower than the benchmarked schemes, since it incorporated the benefits of BiLSTM for reducing the number of inessential frames to be explored.

Table 4 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect to pixel-level AUC based on UMN dataset

Techniques	AUC at the pixel level (in %)
Sparse Coding	89.12
Subgraph Calibration	88.76
Multi-Scale Aggregation	89.54
CNN-BiLSTM	**92.38**

Table 5 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect to pixel-level AUC based on PETS dataset

Techniques	AUC at the pixel level (in %)
Sparse Coding	90.14
Subgraph Calibration	91.28
Multi-Scale Aggregation	90.63
CNN-BiLSTM	**93.29**

Table 6 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect to pixel-level AUC based on UCSD dataset

Techniques	AUC at the pixel level (in %)
Sparse Coding	88.74
Subgraph Calibration	89.12
Multi-Scale Aggregation	90.52
CNN-BiLSTM	**94.18**

Table 7 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect to Detection time incurred per frame (seconds) based on UMN dataset

Techniques	Detection time incurred per frame (seconds)
Sparse Coding	0.0232
Subgraph Calibration	0.0274
Multi-Scale Aggregation	0.0321
CNN-BiLSTM	**0.0165**

Table 8 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect to Detection time incurred per frame (seconds) based on PETS dataset

Techniques	Detection time incurred per frame (seconds)
Sparse Coding	0.0234
Subgraph Calibration	0.0242
Multi-Scale Aggregation	0.0211
CNN-BiLSTM	**0.0182**

Table 9 Performance of the proposed CNN-LSTM abnormal event detection scheme with respect toDetection time incurred per frame (seconds) based on UCSD dataset

Techniques	Detection time incurred per frame (seconds)
Sparse Coding	0.0221
Subgraph Calibration	0.0246
Multi-Scale Aggregation	0.2763
CNN-BiLSTM	**0.0194**

In the third fold of investigation, the proposed CNN-LSTM model is evaluated based on accuracy in abnormal event detection and equal error rate with respect to the datasets of UMN, PETs and USCD. *Figure 4*, *5*, *6* presents the performance of the proposed CNN-LSTM model and the benchmarked abnormal event detection approaches based on detection accuracy with UMN, PETs and USCD datasets. The results confirmed that the proposed CNN-LSTM model is capable in detecting the abnormal events at maximized accuracy, since it derives the merits of temporal and spatial features at the frame and pixel-level for better classification process. Thus, the proposed CNN-LSTM model, independent to the applied datasets confirmed better mean detection accuracy of 12.28%, better than the benchmarked schemes considered to investigation.

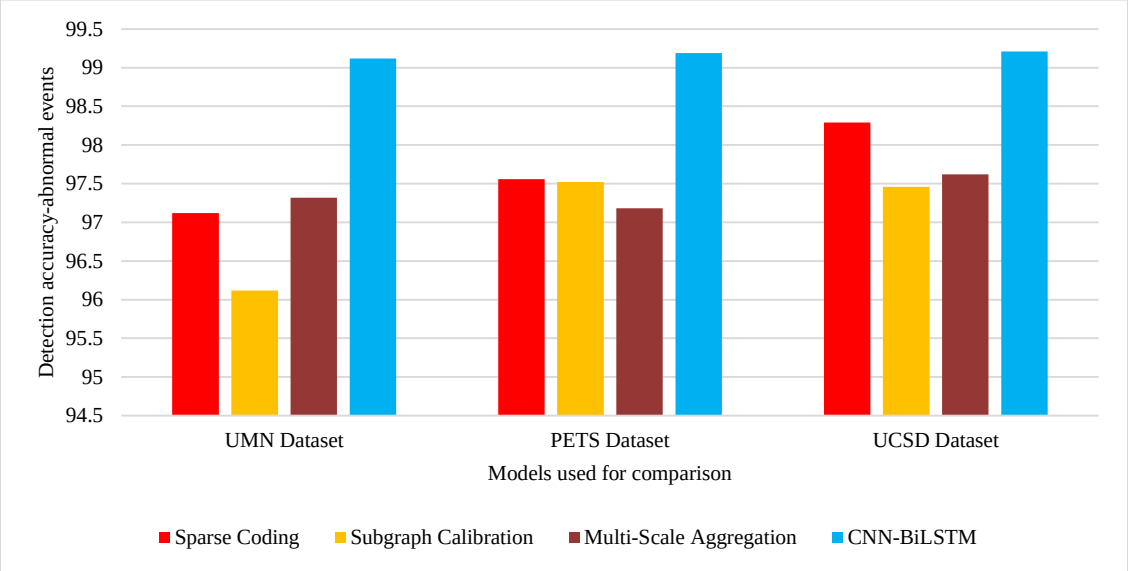

Figure 4 Abnormal event detection-detection accuracy with different datasets

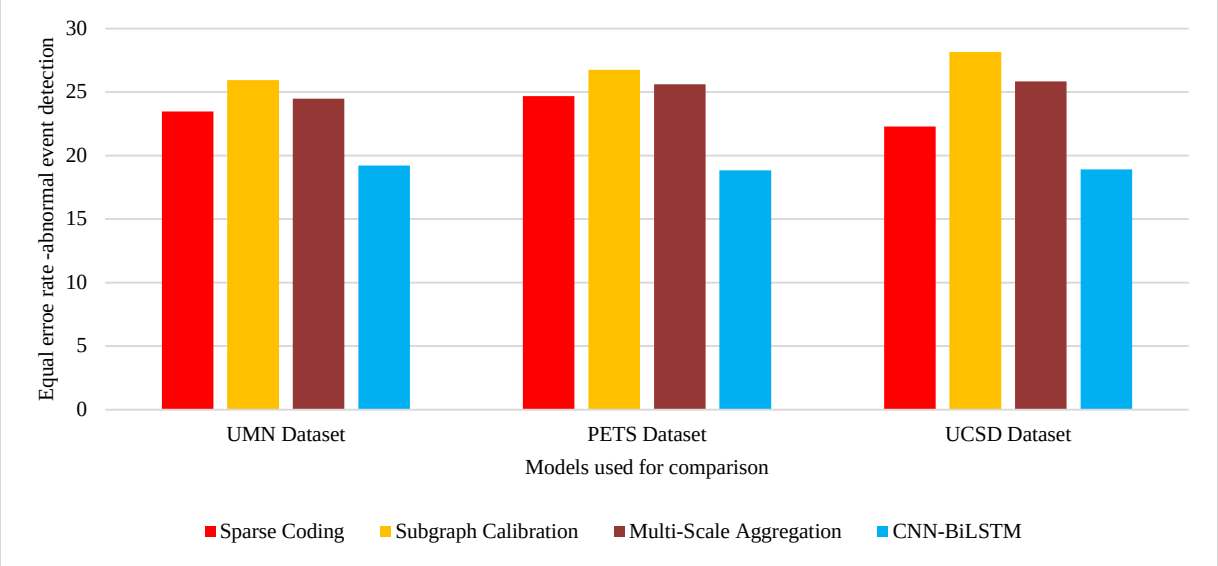

Figure 5 Abnormal event detection-Equal error rate with different datasets

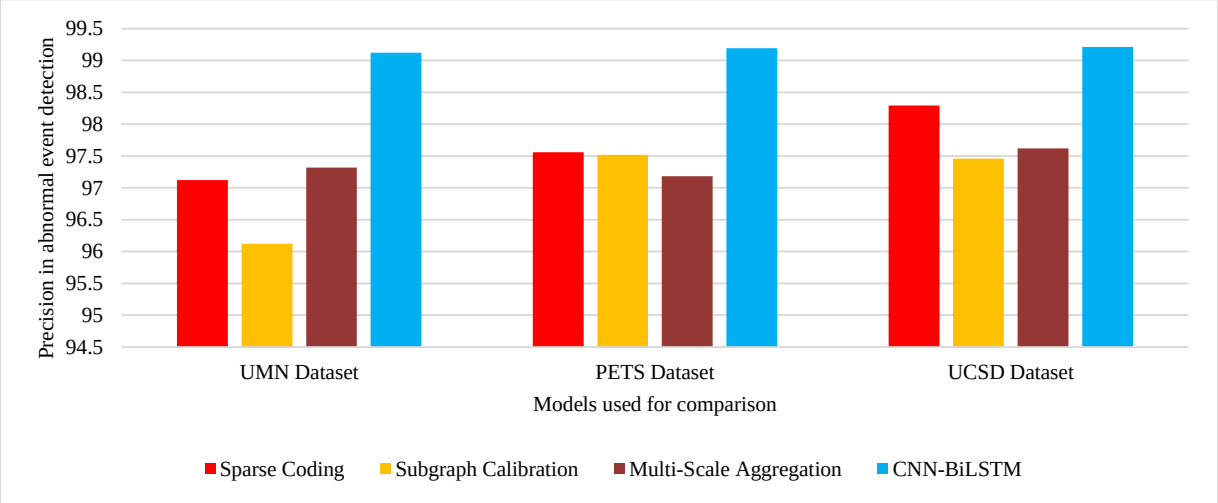

Figure 6 Abnormal event detection-precision with different datasets

Hence, the proposed hierarchical temporal attention-based LSTM encoder-decoder model evaluated based on F-score clearly proved its mean improvement by 12.28%, better than the benchmarked schemes considered to investigation. Finally, the performance of the proposed LSTM encoder-decoder model-based video summarization scheme and the benchmarked approaches are evaluated based on F-score value and the results are presented in *Table 10, 11* and *12*, respectively. Independent to the datasets utilized, the F-score of the proposed LSTM encoder-decoder model-based video summarization scheme is visualized to be always predominant than the benchmarked approaches. This predominant performance of the proposed video summarization model is mainly due to the inclusion of hierarchical temporal attention networks that included global temporal attention and local temporal attention for estimating the relative significance of different key frames with a candidate key frame.

A complete list of abbreviations is shown in *Appendix I.*

Table 10 Performance of the proposed Hierarchical temporal attention-based LSTM encoder-decoder model evaluated based on F-score with respect to UMN dataset

Authors	F-score
Multi video summarization via multi model weighed archetypal analysis [21]	0.9672
Event Retrieval Algorithm using superframe segmentation [22]	0.9718
Proposed	**0.9921**

Table 11 Performance of the proposed Hierarchical temporal attention-based LSTM encoder-decoder model evaluated based on F-score with respect to PETS dataset

Authors	F-score
Multi video summarization via multi model weighed archetypal analysis [21]	0.9621
Event Retrieval Algorithm using superframe segmentation [22]	0.9674
Proposed	**0.9904**

Table 12 Performance of the proposed Hierarchical temporal attention-based LSTM encoder-decoder model evaluated based on F-score with respect to UCSD dataset

Authors	F-score
Multi video summarization via multi model weighed archetypal analysis [21]	0.9732
Event Retrieval Algorithm using superframe segmentation [22]	0.9718
Proposed	**0.9926**

5.Conclusion and future work

In this paper, an abnormal event detection scheme and a video summarization scheme was proposed using the merits of hybrid CNN and bi-directional LSTM and hierarchical temporal attention-based LSTM encoder-decoder model for facilitating proofs to the claims in the insurance sector. The CNN-BiLSTM model was used for abnormal event detection as it facilitates significant extraction of spatio-temporal features from individual frames of video sequences with minimized complexity. Likewise, hand, hierarchical temporal attention-based LSTM encoder-decoder model is incorporated for video summarization, since they are capable of preserving video key information for achieving optimal storage. The experiments of the proposed frameworks conducted using the datasets of PETS, UMN and UCSD, confirmed that the hybrid CNN and bi-directional LSTM model included in the proposed surveillance frame improved the AUC and minimized equal error rate by 4.56% and 7.82%, compared to the state of-the art approaches considered for investigation. The video summarization used in the proposed frameworks was considered to enhance F-score by 9.28% and minimize computational time by 12.39%, better than the baseline video summarization schemes. In the near future, it is also planned to develop a new surveillance framework with the merits of the BMAN and attention based encoder–decoder networks, respectively, to enable the abnormal event and video summarization with the view to compare it with the proposed framework in order their predominance. As a future work, a self-adaptable recognition method for identifying abnormal events through pre-classification techniques can be investigated with different datasets.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

G. Balamurugan: Conceptualization, investigation, writing – original draft, writing – review and editing-data collection, conceptualization, writing – original draft, analysis and interpretation of results. **J. Jayabharathy:** Study conception, design, data collection, supervision, investigation on challenges and draft manuscript preparation.

References

[1] Li B, Leroux S, Simoens P. Decoupled appearance and motion learning for efficient anomaly detection in surveillance video. Computer Vision and Image Understanding. 2021.

[2] Gaikwad B, Karmakar A. Smart surveillance system for real-time multi-person multi-camera tracking at the edge. Journal of Real-Time Image Processing. 2021; 18(6):1993-2007.

[3] Kalair K, Connaughton C. Anomaly detection and classification in traffic flow data from fluctuations in

the flow–density relationship. Transportation Research Part C: Emerging Technologies. 2021.
[4] Kolekar S, Gite S, Pradhan B, Kotecha K. Behavior prediction of traffic actors for intelligent vehicle using artificial intelligence techniques: a review. IEEE Access. 2021; 9:135034-58.
[5] Meng Q, Shang B, Liu Y, Guo H, Zhao X. Intelligent vehicles trajectory prediction with spatial and temporal attention mechanism. IFAC-PapersOnLine. 2021; 54(10):454-9.
[6] Chu W, Wuniri Q, Du X, Xiong Q, Huang T, Li K. Cloud control system architectures, technologies and applications on intelligent and connected vehicles: a review. Chinese Journal of Mechanical Engineering. 2021; 34(1):1-23.
[7] Nittayasoot N, Peterson AB, Thammawijaya P, Parker EM, Sathawornwiwat A, Boonthanapat N, et al. Evaluation of a hospital-based injury surveillance system for monitoring road traffic deaths in Phuket, Thailand. Traffic Injury Prevention.2019; 20(4):365-71.
[8] Balamurugan G, Jayabharathy J. A comparative analysis of event detection and video summarization. In ambient communications and computer systems 2022 (pp. 577-86). Springer, Singapore.
[9] Fu H, Wu L, Jian M, Yang Y, Wang X. MF-SORT: simple online and realtime tracking with motion features. In international conference on image and graphics 2019 (pp. 157-68). Springer, Cham.
[10] Zhu J, Zhang S, Yang J. Online multi-object tracking using single object tracker and Markov clustering. In international conference on image and graphics 2019 (pp. 555-67). Springer, Cham.
[11] Borji A, Cheng MM, Hou Q, Jiang H, Li J. Salient object detection: a survey. Computational Visual Media. 2019; 5(2):117-50.
[12] Fan DP, Wang W, Cheng MM, Shen J. Shifting more attention to video salient object detection. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 8554-64).
[13] Santamaria AF, Raimondo P, Tropea M, De RF, Aiello C. An IoT surveillance system based on a decentralised architecture. Sensors. 2019; 19(6):1-23.
[14] Li R, Pereira FC, Ben-akiva ME. Overview of traffic incident duration analysis and prediction. European Transport Research Review. 2018; 10(2):1-13.
[15] https://www.v7labs.com/blog/object-detection-guide. Accessed 27 August 2022.
[16] Benterki A, Boukhnifer M, Judalet V, Maaoui C. Artificial intelligence for vehicle behavior anticipation: hybrid approach based on maneuver classification and trajectory prediction. IEEE Access. 2020; 8:56992-7002.
[17] Al-omari A, Shatnawi N, Khedaywi T, Miqdady T. Prediction of traffic accidents hot spots using fuzzy logic and GIS. Applied Geomatics. 2020; 12(2):149-61.
[18] Shen J, Tao D, Li X. Modality mixture projections for semantic video event detection. IEEE Transactions on Circuits and Systems for Video Technology. 2008; 18(11):1587-96.
[19] Yi D, Su J, Liu C, Chen WH. Trajectory clustering aided personalized driver intention prediction for intelligent vehicles. IEEE Transactions on Industrial Informatics. 2018; 15(6):3693-702.
[20] Martinesco A, Netto M, Neto AM, Etgens VH. A note on accidents involving autonomous vehicles: interdependence of event data recorder, human-vehicle cooperation and legal aspects. IFAC-PapersOnLine. 2019; 51(34):407-10.
[21] Ji Z, Zhang Y, Pang Y, Li X, Pan J. Multi-video summarization with query-dependent weighted archetypal analysis. Neurocomputing. 2019; 332:406-16.
[22] Wan S, Xu X, Wang T, Gu Z. An intelligent video analysis method for abnormal event detection in intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems. 2020; 22(7):4487-95.
[23] Gelmini S, Strada S, Tanelli M, Savaresi S, De TC. Automatic crash detection system for two-wheeled vehicles: design and experimental validation. IFAC-PapersOnLine. 2019; 52(5):498-503.
[24] Jahangiri A, Berardi VJ, Machiani SG. Application of real field connected vehicle data for aggressive driving identification on horizontal curves. IEEE Transactions on Intelligent Transportation Systems. 2017; 19(7):2316-24.
[25] Balamurugan G, Jayabharathy J. A study on moving object recognition for video surveillance applications. Journal of Advanced Research in Dynamical and Control Systems. 2019:157-67.
[26] Lei J, Luan Q, Song X, Liu X, Tao D, Song M. Action parsing-driven video summarization based on reinforcement learning. IEEE Transactions on Circuits and Systems for Video Technology. 2018; 29(7):2126-37.
[27] Yuan Y, Mei T, Cui P, Zhu W. Video summarization by learning deep side semantic embedding. IEEE Transactions on Circuits and Systems for Video Technology. 2017; 29(1):226-37.
[28] Kalaivani P, Roomi SM, Jaishree B. Video event representation for abnormal event detection. In international conference on circuits and systems 2017 (pp. 463-8). IEEE.
[29] Ahmed SA, Dogra DP, Kar S, Roy PP. Trajectory-based surveillance analysis: a survey. IEEE Transactions on Circuits and Systems for Video Technology. 2018; 29(7):1985-97.
[30] Moses TM, Balachandran K. A classified study on semantic analysis of video summarization. In international conference on algorithms, methodology, models and applications in emerging technologies 2017 (pp. 1-6). IEEE.
[31] Wang T, Miao Z, Chen Y, Zhou Y, Shan G, Snoussi H. Aed-net: an abnormal event detection network. Engineering. 2019; 5(5):930-9.
[32] Chu W, Xue H, Yao C, Cai D. Sparse coding guided spatiotemporal feature learning for abnormal event

detection in large videos. IEEE Transactions on Multimedia. 2018; 21(1):246-55.
[33] Kalaivani P, Roomi SM. Towards comprehensive understanding of event detection and video summarization approaches. In second international conference on recent trends and challenges in computational models 2017 (pp. 61-6). IEEE.
[34] Meghdadi AH, Irani P. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. IEEE Transactions on Visualization and Computer Graphics. 2013; 19(12):2119-28.
[35] Tejero-de-pablos A, Nakashima Y, Sato T, Yokoya N, Linna M, Rahtu E. Summarization of user-generated sports video by using deep action recognition features. IEEE Transactions on Multimedia. 2018; 20(8):2000-11.
[36] Coşar S, Donatiello G, Bogorny V, Garate C, Alvares LO, Brémond F. Toward abnormal trajectory and event detection in video surveillance. IEEE Transactions on Circuits and Systems for Video Technology. 2016; 27(3):683-95.
[37] Muhammad K, Hussain T, Del SJ, Palade V, De AVH. DeepReS: a deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios. IEEE Transactions on Industrial Informatics. 2019; 16(9):5938-47.
[38] Kong L, Dai R. Object-detection-based video compression for wireless surveillance systems. IEEE MultiMedia. 2017; 24(2):76-85.
[39] Ye O, Deng J, Yu Z, Liu T, Dong L. Abnormal event detection via feature expectation subgraph calibrating classification in video surveillance scenes. IEEE Access. 2020; 8:97564-75.
[40] Kumar K, Shrimankar DD. F-DES: fast and deep event summarization. IEEE Transactions on Multimedia. 2017; 20(2):323-34.
[41] Lee S, Kim HG, Ro YM. BMAN: bidirectional multi-scale aggregation networks for abnormal event detection. IEEE Transactions on Image Processing. 2019; 29:2395-408.
[42] Thomas SS, Gupta S, Subramanian VK. Perceptual video summarization—a new framework for video summarization. IEEE Transactions on Circuits and Systems for Video Technology. 2016; 27(8):1790-802.
[43] 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. 2019; 30(6):1709-17.
[44] Pinho RR, Tavares JM, Correia MV. An improved management model for tracking missing features in computer vision long image sequences. Article in International Scientific Journal. 2006.
[45] Song H, Sun C, Wu X, Chen M, Jia Y. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. IEEE Transactions on Multimedia. 2019; 22(8):2138-48.
[46] Yan S, Smith JS, Lu W, Zhang B. Abnormal event detection from videos using a two-stream recurrent variational autoencoder. IEEE Transactions on Cognitive and Developmental Systems. 2018; 12(1):30-42.
[47] Huang C, Wang H. A novel key-frames selection framework for comprehensive video summarization. IEEE Transactions on Circuits and Systems for Video Technology. 2019; 30(2):577-89.
[48] Dilawari A, Khan MU. ASoVS: abstractive summarization of video sequences. IEEE Access. 2019; 7:29253-63.

G. Balamurugan completed Bachelor degree [B.Tech., Information Technology] in 2013 under Pondicherry University and Master Degree [M.Tech., Computer Science & Engineering] in 2015 under Pondicherry University. He secured University second rank in Master Degree under Pondicherry University. He holding a Teaching Experience of 7 years and currently working as an Assistant Professor in Department of Computing Technologies at SRM Institute of Science and Technology. Holding a life time membership in IAENG. Research domain is Video Processing and Information Security. Published various Research papers in reputed Journals and participated in IEEE, Springer Conferences. His area of interest includes Artificial Intelligence and Information Security.
Email: gbalamurugan1991@pec.edu

J. Jayabharathy is an Associate Professor in Department of Computer Science and Engineering at Puducherry Technological University, Puducherry. She received her Ph.D. in CSE in the year 2014. He received his M.Tech., from Pondicherry University in the year 1999 and B.Tech., in the year 1998. She has published more than 60 papers in international journals and conferences. Her area of interest includes Distributed Computing, Text Mining and Social Area Networks.
Email: bharathyraja@pec.edu

Appendix I

S. No.	Abbreviation	Description
1	AUC	Area Under the ROC Curve
2	Bi-LSTM	Bidirectional Long Short-Term Memory
3	BMAN	Bidirectional multi-scale aggregation networks
4	CNN	Convolutional Neural Network
5	HOG	Histogram of Oriented Gradients
6	PETS	Performance Evaluation of Tracking and Surveillance
7	ROC	Receiver Operating Characteristic Curve
8	SURF	Speeded Up Robust Features
9	SIFT	Scale Invariant Feature Transform
10	UMN	Unusual crowd Activity Dataset of University of Minnesota
11	UCSD	University of California San Diego