

Original Article

Bi-Attention LSTM with CNN based Multi-task Human Activity Detection in Video Surveillance

Shankargoud Patil¹, Kappargaon S. Prabhushetty²

¹Assistant Professor, Dept. of ECE, S. G. Balekundri Institute of Technology, Belagavi, Karnataka, India

²Adjunct faculty, Dept. of ECE, VeerappaNisty Engineering College, Hasanapur, Shorapur, Yadgir, Karnataka, India

¹shankar.a.patil@gmail.com, ²kprabhushetty1@gmail.com

Abstract — Computer vision and pattern recognition, the hot subjects include crowd analysis and anomalous trajectories detection. Anomaly detection is a technique for distinguishing between different patterns and identifying uncommon patterns in a short amount of time. Abnormal event detection and localization is a difficult research challenge due to its complexity. It's made to detect unusual events in monitoring videos automatically. In the proposed method, humans' normal and abnormal activities are detected through Deep Learning (DL) and image processing. To use the proposed Bi-Attention Long Short-Term Memory (Bi-Attention LSTM) model to extract just the necessary spatial and temporal information from videos and to predict the multi-task activities of humans as abnormal or normal using the introduced Convolutional Neural Network (CNN). Video is taken as input and is then transformed into frames, and background subtraction is used to identify the moving objects (people) in the video frame. The proposed Convolutional Neural Network (CNN) with Bi-Attention LSTM model extracts temporal and spatial characteristics before classifying to determine if a specific human action is normal or abnormal. In terms of accuracy, sensitivity, specificity, error, precision, $F1_score$, FPR , $kappa$, and MCC , a performance analysis compares the proposed system to the existing system. On the UMN dataset, Area under the Curve (AUC) and Equal Error Rate (EER) are also considered for comparison. The proposed method finds human operations the most sensible, giving 98.4436% accuracy, which is higher than other existing methods. The results explore the efficacy of the proposed system for classifying human activities from the videos.

Keywords — Bi-Attention Long Short-Term Memory, Convolutional Neural Network, Human Activity Detection, Normal and Abnormal Activities, and Video Surveillance.

I. INTRODUCTION

Detecting suspicious activities is significant from a surveillance video to avert theft, terrorism, illegal parking, accidents, vandalism, chain snatching, fighting, crime, and other suspicious activities. It is hard to monitor all the places continuously. As a result, automatic video surveillance is needed that can monitor the activities of humans in real-time,

followed by classifying them as unusual and usual activities. The study [1] reviews the present status of the video surveillance systems that intend to find effective practices for image and video processing. It also spots the research issues for the subsequently generated systems. For the past five decades, most of the study has been attempting to replace an operator's work with a video processing algorithm. This will have the ability to perform specific tasks, thereby reducing the burden of manpower. Though more efforts have been put into the operation, it has retrieved a few brilliant and innovative algorithms, and these methods can only perform small tasks like object detection and face recognition. The accuracy of the systems is typically far from fulfilling when these scene conditions have not been perfect.

Developments in research during the early decade have rejuvenated the potentials for spontaneous surveillance systems. Specifically, these developments involve Deep Learning (DL), Machine Learning (ML), and Distributed Computational infrastructure, namely fog computing, edge computing, and cloud computing. These integrated techniques have been expected to enhance the surveillance algorithm's accuracy and introduce smart analytics, thereby minimizing the system's response time to generate meaningful alarms [2]. Similarly, the article [3] introduced a Hierarchical Spatio-Temporal Model (HSTM) for interactional activity and one-person action identification by modeling temporal and spatial constraints simultaneously. Empirical outcomes exhibit significant enhancements on discriminative power and capability. Detection of humans is the essential stage in recognizing activities. Hence, the study used a Convolutional Neural Network (CNN) model for human detection [4]. The technique can assist the law enforcement supports to detect abnormal activities from video surveillance like accidents, criminal activities (stealing, fighting, and so on), and activities that result in property damage [5]. Although it might produce false alarms as the present version has been trained on openly available data for multiple activities, it will be complex to directly execute it in the video surveillance settings [6].

Likewise, the article [7] introduced a system for handling the abnormal behavior of humans in a video. A rule-based strategy is utilized to distinguish the activities as abnormal or normal. A few more features must be executed by



concentrating on a few more abnormal activities, and numerous overlapping objects must be considered. In the same way, the study [8] introduced a framework for detecting abnormal activities, which comprises dynamics and appearance in addition to geometric associations among several entity interactions in any video activity [9]. The usage of graph kernel to find the resemblance amongst two graphs affords robustness to small distortions to the structure due to noise in the data. The empirical outcomes performed better than traditional methods, namely bag-of-visual words, dense trajectories, etc., confirming the proposed system's efficiency [10]. On the contrary, the proposed work aims to detect the multi-task activities of humans from video and classify it as normal or abnormal. For this purpose, it introduces Convolutional Neural Network (CNN) for feature extraction (spatial and temporal features). It then performs classification by the proposed trained Bi-Attention Long Short-Term Memory (Bi-Attention LSTM) model.

The major contributions of this study are listed below.

- An effective model Bi-attention LSTM with CNN is used to detect normal and abnormal behavior of human beings in a congested area.
- Video clips of the surveillance are segmented and remove the unwanted things, extracting the spatial and temporal features by using the proposed Convolutional Neural Network (CNN).
- To prognosticate the behavioral events by the introduced Bi-Attention Long Short-Term Memory (Bi-Attention LSTM) model.
- To analyze the performance of the introduced methodology in terms of significant standard performance metrics to evaluate its efficiency in detecting the normal and abnormal activities of humans from video surveillance.
- The proposed technique is very sensitive to detecting human activities from video input, giving 98.4436% accuracy. Furthermore, the research will compare the existing and proposed systems using the UMN dataset.

The rest of the paper is organized as section II explores various methodologies used by the state-of-the-art methods. Following this, the overall proposed system, along with the process involved in this context, are explained in section III. The results obtained after the implementation of the introduced methods are explored in section IV. Finally, an overall summarization of the planned method is presented along with the future work in section V.

II. REVIEW OF EXISTING WORK

Detection of normal and abnormal activities is crucial because it avoids abnormal behaviors such as theft, fighting, etc. In the olden days, manual detection methods were used and did not provide an accurate result due to limitations like difficulty monitoring several displays. After the tremendous growth of computer technology, a computer-aided automatic detection method was developed to detect normal and

abnormal activities. These methods are more helpful for avoiding abnormal human behaviors and securing a human being's life.

The existing system has used numerous methods for detecting human activities from video surveillance. These different kinds of approaches are discussed in this section.

A. Literature survey-based on HAR without Deep learning

The suspicious activity of humans can be detected from video surveillance which has transformed into a burgeoning area for research in computer vision and video processing. It also involves Human Activity Recognition (HAR) and classifies them into abnormal and normal activities. In recent years, the usage of video surveillance has been enhancing each day to monitor all human activities to prevent humans' suspicious activities

Malek Al-Nawashi et al. [11] had introduced a video relied surveillance system that is automatic and of real-time. At the same time, it can perform learning of semantic scenes, detect abnormal actions, and track in any academic ambiance. This particular system has been developed by partitioning into three stages, namely pre-processing stage, abnormal HAR stage, and content relying on image retrieval (IR) stage. At last, the proposed method has been executed by using MATLAB tools. The simulation outcomes of the study revealed the efficiency of the introduced system. It is effective with respect to performance and accuracy, thereby exhibiting a minimum false alarm rate. However, this method does not detect a small fire covered by dark smoke, and the cost of the system was slightly higher due to the sensor connection. In addition, the method's effectiveness was majorly affected by the recognition of the zone describing the human body in the case of nearby persons.

Rajesh Kumar Tripathi et al. [12] had presented various abnormal activities like theft detection, wild object detection, fall detection, prohibited parking detection, fire detection, accidents, and detection of violent activity. This study has aimed to analyze six different suspicious HAR with their common structure to the investigators of this area. Numerous operations have been undertaken to prevent violent activities like fighting, vandalism, punching, shooting, and hitting. A static video camera of single-use has been utilized for violence activity detection. But, at times, the introduced system does not fit properly in handling occlusions.

Consequently, a system with multi-view features was introduced by a few other investigators to solve this issue. However, it needs significant contribution amongst every view at the minimum step-level to detect abnormal activity. In the future, the study incorporates high enhancement with respect to frame rate, reduction in false alarms, and accuracy. This can also be improvised to detect small fires covered by dark smoke in the far distance. A video surveillance system has different levels.

Amira Ben Mabrouk and Ezzeddine Zagrouba[13] suggested a level of behavior modeling and behavior representation. The feature choice that has been utilized to illustrate the

mobile object is a complex task as it impacts the description and the examination of this behavior significantly. Hence, it is vital to choose robust features to scene renovations like rotation, cluttered backgrounds, occlusion, etc. It is also sensitive at a low rate to the alterations in the object's appearance to find the discriminative and relevant information regarding the object's moving behavior. But, owing to the variation of atmospheric conditions in which video surveillance systems need to locate and monitor the object, the systems must be active. In these circumstances, the smallest margin error of the systems is expected.

B. Literature survey-based on HAR with Deep learning

For instance, running is typically a day-to-day activity most often people perform. On the other hand, when any person is running on any road, it is considered an abnormal behavior that needs to be prompted. The drawbacks as mentioned above have to be overcome. For this purpose, the introduced systems utilized large training data that comprise all the probable circumstances. Examples might include employing Deep Learning (DL) to work on large datasets effectively. The deep architectures of DL have made its usage highly grow to attain a high learning capacity.

Chhavi Dhiman and Dinesh Kumar Vishwakarma [14] had presented a traditional abnormal HAR and use of deep approaches along with different kinds of information accessible like two or three-dimensional data. This study affords designing features corresponding to abnormal HAR in a video according to the application or context like Ambient Assistive Living (AAL), fall detection, surveillance, indoor security, or group analysis through depth, RGB, and skeletal indication. However, the algorithm was more complicated to detect the activity. Also, sometimes the binary silhouette fails to describe the pose if it occurs spontaneously. Rapid variation of speed and direction causes false-positive rate error. In addition, the method was more sensitive to the parameters of the control point in the trajectory.

Hironori Hattori et al. [15] had introduced a strategy that considered the viewpoint of pedestrian's projection on an image plane. This study also designed the appearance of pedestrians underneath the Synthetic object Occlusion (SOC). These training methods based on synthesis are highly fit for the present model of Data-Hungry Object Detectors (D-HODs). However, the study concentrated on using geometry for the synthesis process. This is the initial step in enhancing the knowledge for the process of synthesis. But, both localization and identification of tasks were mostly interdependent methods. And the effectiveness of the detection process was highly affected by the process of pose assessment. And the high-level semantic discernment of the scene, which was used to create a wide range of human representations, has yet to be investigated.

Rashmika Nawaratne et al. [16] introduced Spatio-temporal anomaly detection using Active Learning (AL) and DL for video surveillance in real-time. The presented model relies

on a Spatio-temporal autoencoder that comprises several convolutional layers and ConvLSTM layers. The convolutional layers of the autoencoder model learn spatial regularities. The ConvLSTM of the autoencoder model learns the temporal regularities conserving the video stream's spatial structure. Experiments have been carried out on benchmark datasets, and the outcomes explored robustness, accuracy, contextual indicators, and minimum computational overhead of the introduced strategy authorizing its wide usage in urban and industrial settings. However, knowledge was required before designing good features, and it takes time to extract them, making it impossible to detect real-time anomalies. On the other hand, detection and localization can be parallelized to reduce run time and increase FPS due to the series implemented by ISTL.

Ahmad Jalal et al.[17] had presented an automatic video surveillance system having minimized False Negative (FN) detection. But, this database was difficult because most in-depth positions have similar appearances to various processes, so it was not suitable for real-time application.

Karishma Pawar and Vahida Attar [18] had reviewed the DL methodologies for suspicious activity detection from video surveillance. As a result of this research, a graphical taxonomy has been suggested that relies on the anomaly types, level of detecting anomalies, and measuring anomalies for detecting anomalous activity. However, Multi-view anomaly detection, on the other hand, has received less attention. It is important to find a balance between real-time processing and the right level of accuracy. Detecting camouflage is also a difficult task.

Wei Huang et al. [19] presented a localizing the moving humans in the videos emphasized individual frames. Quantitative and qualitative analysis has been undertaken from a statistical viewpoint to assess all the localizing results that rely on two general computations. The proposed technique is explored to perform better than other traditional systems. Sophisticated DL-based models have to be examined and proposed in the near future [20]. But, the shallower network's generalization capability will always decline, and a deeper network structure is always chosen for its attractive potential in reflecting more intricate non-linear relationships and ensuring more promising generalization capability for solving a degeneration problem in the network. Hongxun Yao [21] presented extensive accessibility of cameras, and the developing requirement to save the public has altered the researcher's attention in video surveillance from individual behavior analysis to group behavior analysis in the multi-camera networks. This study includes all the recent developments in HAR from video surveillance. It mainly focussed on localizing a person and re-identification, tracking a person, crowd behavior detection and recognition, learning the abnormal behavior patterns followed by detection and action clustering in addition to recognition.

Tian Wang et al. [22] had introduced a method for abnormal event detection that relies on video analysis. This method involves the Movement Feature Descriptor (MFD) and

classification technique. MFD encodes all the movement information that relies on examining the ideal flow corresponding to the area of interest has also been proposed. Subsequently, the Hidden Markov Model (HMM) has been derived from differentiating the hidden states through the MFD analysis with the Probability Property (PP). The PP of the histogram relying on MFD has also been examined. Therefore, the proposed model fits to be applied for abnormal classification. However, this method was not suitable for the multi-class problem. The system's accuracy is based on the predicted model; to some extent, the forecast outcome of these frames can be treated as noise. These noises were filtered to find accuracy, so the method was time-consuming and unsuitable for real application.

Artur Jordao et al. [23] suggested a new ConvNet architecture to exhibit the patterns via the network layers. The proposed method is a simple way to enhance recognizing the activities rather than introducing highly complicated architectures. This will serve as a future direction to construct the ConvNets architecture. The empirical outcomes explored that the introduced techniques accomplish notable enhancements and perform well than the existing techniques. Yet, ConvNets architecture was not compatible because they recorded a short transient format and were vulnerable to noise during data acquisition due to the convolutional kernel design.

Shaohua Wan et al. [24] presented an exploration of two varied HAR datasets like Pamap2 and UCI. It uses various algorithm models like Long Short-Term Memory (LSTM), Multi-Layer Perceptron (MLP), Bi-LSTM (Bi-directional Long Short-Term Memory), Convolutional Neural Network (CNN), and Support Vector Machine (SVM) for an individual dataset. Experiments have been carried out using four Neural Network (NN) models' structures [25]. Effective outcomes have been attained for HAR. However, the four neural network models utilized in the experiment have a weak structure; in addition, the activities of some persons cannot be accurately identified.

Fatma Najar et al. [26] had presented a Fixed Point Estimation (FPE) method developed to learn the multi-variate and Gaussian model that uses a complete co-variance matrix. The proposed methodology has been employed with an Expectation-Maximization (EM) algorithm to construct an unsupervised learning method for HAR. Empirical outcomes exhibited the proposed system's efficacy. Yet, this introduced method has to be extended to online settings. This might enhance the model's learning [27]. Technology has also been presented to detect suspicious activities that might occur during any close communications between people. However, this procedure was lengthy, complicated, non-reproducible, and very subjective. And the symmetry around the mean and the rigidity of the Gaussian density were limited.

Kwang-Eun Ko and Kwee-Bo Sim [28] explore technology to detect this kind of abnormal behavior. The main characteristics of this research assume that the behavior of

humans is comprised of a structure of fixed appearances and their temporal associations subsidized to HAR [29]. To enhance the results' accuracy in recognizing the static pose, a DL model has been utilized for object identification in real-time and is named the YOLO network, which recognizes and then differentiates the activities in the video. The overall outcomes explored the possibility and potentiality of the introduced technique [30].

Joey Tianyi Zhou et al. [31] suggested a new neurological network (called anomalinet) for detecting anomalies by in-depth attainment of feature learning, rare representation, and dictionary learning in three collective neurological processing modules. Yet, many unusual occurrences of jogging happen in the background where most normal walking takes place.

The literature, as mentioned above reviews contains several drawbacks such as does not detect a small fire covered by dark smoke, the cost of the system being slightly higher, being affected by atmospheric conditions, weak network structure, and difficult to predict the class. To overcome these issues mentioned in previous methods, this paper presents a mechanism with CNN integrated Bi-Attention LSTM.

III. PROPOSED METHODOLOGY

Automatic video surveillance is mostly used for detecting abnormal activities recognition in a crowded place. Abnormal events in public places like sudden violence have extremely precious public welfare and have been a hard-to-resolve problem for social authority. Smart City is demanding full attention of surveillance cameras in the meeting area to address these social issues. Due to human inattention and exhaustion, the duty of watching many monitors by security officers gets increasingly difficult as the number of surveillance cameras grows. Aside from that, strange events are rare and occur seldom. Due to a variety of factors, such monitoring is frequently ineffective. This complicates the oversight role. Therefore, the demand for an intelligent video surveillance system that automatically detects abnormal behavior is increasing. In the proposed method, Convolutional Neural Network (CNN) and Bi-Attention Long Short-Term Memory (Bi-Attention LSTM) are used to extract features and predict the events in the video based on the activities of humans by classifying them as normal or abnormal automatically. The video is taken as input collected from CCTV footage, and these video files are converted into frames and subtracting the background. Next, data is processed for feature extraction. While extracting feature information, CNN integrates with Bi-Attention LSTM because CNN can extract local and global properties into one database. Bi-attention LSTM introduces a simultaneous two-observation mechanism to predict normal and suspicious actions of humans. Fig. 1 shows the proposed method's working process.

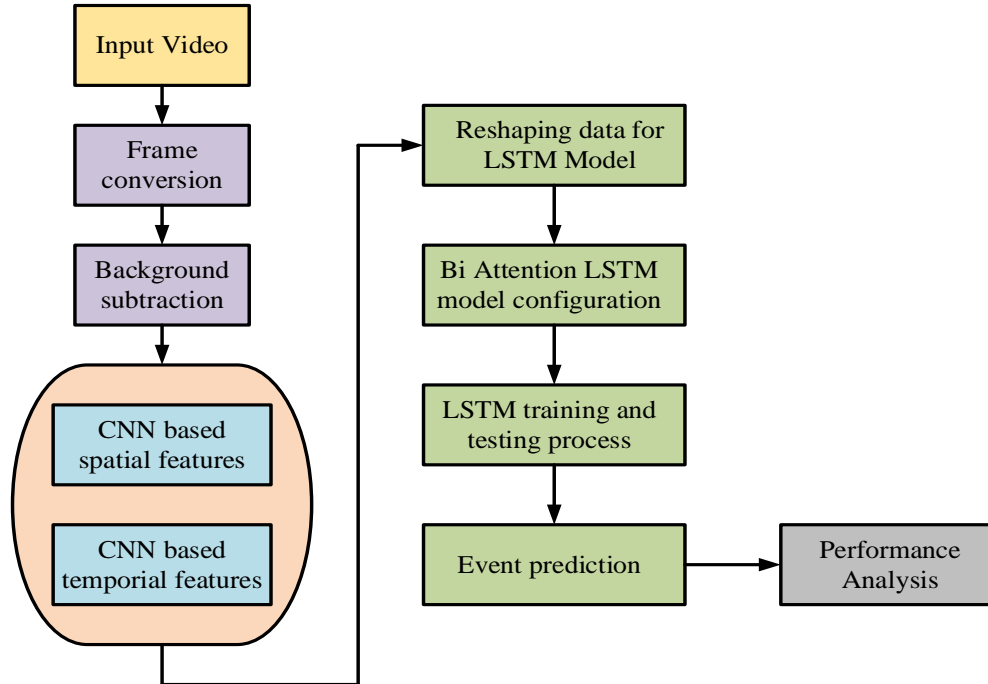


Fig.1. Overall view of the proposed system for event detection from video

At first, the video is taken as input. This video is then converted into frames. For a video, frame per rate is calculated. If there are five frames per rate, a particular frame (1st or 5th frame) is converted into an image. So, the extent to which the frame per rate is set as minimum generates more images. Thus, the video is converted into images based on the frame per rate. Following this, background subtraction is performed, an extensively utilized methodology to detect the moving objects from the video. After this detection, spatial and temporal features are extracted using the proposed CNN. Here, spatial features consist of spatial or location information.

On the other hand, temporal features comprise any feature associated with date or time. After extracting both of these features, it is fed into the LSTM model. In this step, the proposed Bi-Attention LSTM model is configured. Subsequently, the training and testing process is carried out for the proposed Bi-Attention LSTM model. This trained model helps in event prediction (to find humans' abnormal and normal activity from the video). Finally, the performance of the overall proposed system is analyzed by considering specific significant metrics to determine the effectiveness and efficiency of this system for detecting human activities from videos.

A. Background Subtraction

The defining features of this process are how it outlines and updates the corresponding background model. $F_t^{ch}(m,n)$ and $B_t^{ch}(m,n)$ are utilized to indicate the value of the channel ch corresponding to the pixel at a specific location (m,n) and at time t for the respective video stream taken as input in

addition to the background model. When the spatial location is not relevant, the location of the pixel (m,n) is dropped.

a) Foreground detection

Pixels of the video frame taken as input that the background model could not adequately explicate is considered a foreground object. Methods that are deficient in a statistical framework categorize the latest pixel from the foreground.

$$|F_t^{ch}(m,n) - B_t^{ch}(m,n)| > Th \quad (1)$$

where Th indicates the threshold defined by the user. Moreover, the main drawback of this strategy is that it uses only one threshold for every pixel model, though few pixels might explore many variations compared to others. In accordance with this, techniques that afford a variance measure for an individual pixel are desirable. Accordingly, methods that model pixels as a Probability Density Function (PDF) categorize the latest pixel from the foreground.

$$p(F_t^{ch}|B_t^{ch}) < Th^{ch} = \delta\Psi^{ch} \quad (2)$$

where ch is a channel. In addition, the threshold Th^{ch} is fixed proportional to the assessed variation, Ψ^{ch} . To confirm that a pixel is categorized as from foreground based on a specific condition. This is true only when the pixel is exterior to the usually perceived variance level.

B. Feature Extraction - Convolutional Neural Network (CNN)

The proposed Convolutional Neural Network (CNN) based on DL is an efficient method for image processing due to peculiar kinds of pooling and convolutional layers. Due to the deep nature of the network, the gradient or input information passing via many layers disappears over time

when it reaches the last layer of the network. SpatialTemporalNet solves these gradient disappearance issues by connecting each layer with the identical size of features with one another. The main idea of using this

architecture for feature extraction is that many genetic features can be attained when the network is deeper. The architecture of SpatialTemporalNet is shown in Fig. 2.

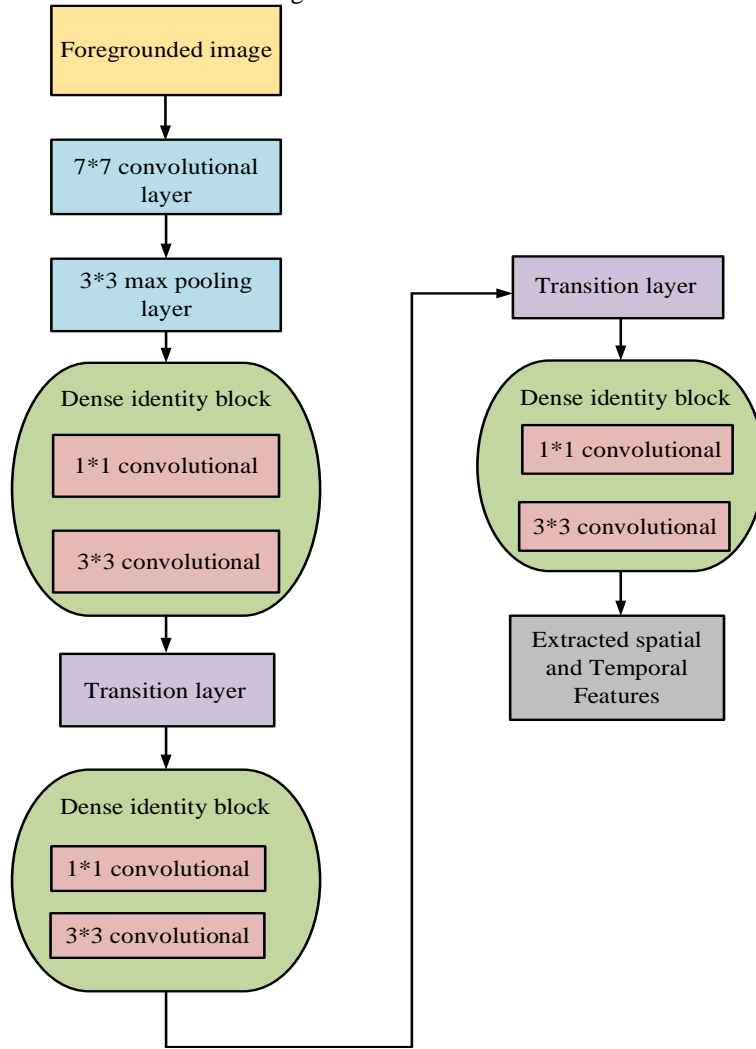


Fig.2. Architecture of SpatialTemporalNet

The above architecture consists of a pooling and convolutional layer, three transition layers, and four dense blocks. After this, the last layer, that is, the classification layer exists. The initial convolutional layer undertakes 7*7 convolutions and max-pooling (3*3). Subsequently, this network comprises a dense block and three individual sets comprising a transition layer and a dense block. The formulae for calculating the volume layer and the downsampling layer are given below. The convolution layer formula and Downsampling layer formula are expressed as,

$$x_j^l = f \left(\sum_{i \in M_j} x_i^{l-1} k_{ij}^l + b_j^l \right) \quad (3)$$

$$x_j^l = f \left(\frac{1}{n} \sum_{i \in M_j} x_i^{l-1} + b_j^l \right) \quad (4)$$

Denseconnectivity corresponding to the SpatialTemporalNet is attained by fetching connections directly from any network layer to any other network layer. The last network layer attains the feature maps of preceding layers. Hence, upgrading the gradient flow all over the network. This needs the feature map concatenation of the previous layers. This could not be accomplished if all of the feature maps have equal sizes; however, as the proposed CNN aims to downsample the feature map size, the SpatialTemporalNet architecture is partitioned into multiple and dense blocks connected, as outlined in Fig. 2. Between these blocks are the layers indicated as transition layers, and these layers comprise a normalization layer, a convolution layer (1*1),

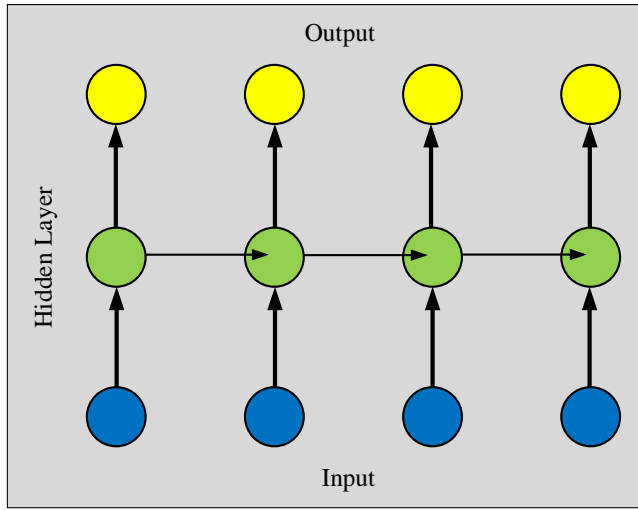
and an average pooling layer (2*2). A feature extractor is included into this backbone network. Object detection networks were mostly used as backbone networks, with no categorization blocks. To converge quickly on thermal videos, the weights of these backbone networks are initialized with ImageNet pre-trained values.

$$f_0 = GAP(B(x(t))) \quad (5)$$

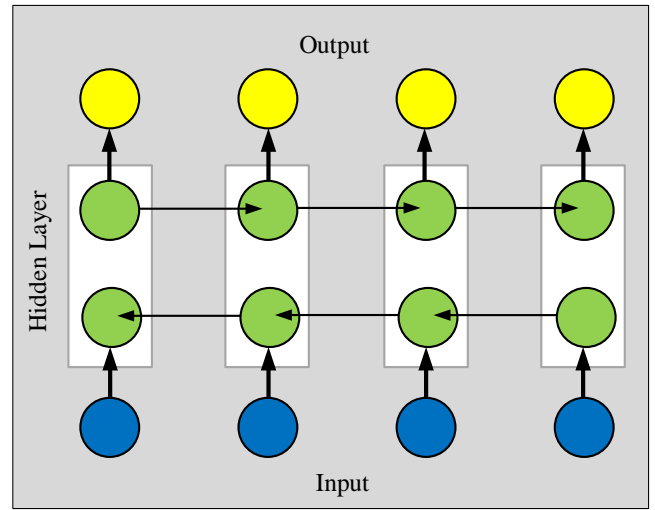
where, f_0 is the output feature vector, GAP is global average pooling operation, $B(\cdot)$ stands for the backbone network, and $x(t)$ is non-linearity. The feature extraction process from the proposed model is employed on the individual network layers other than the last classification layer. The overall representation of spatial and temporal features obtained is interpreted as a dimension vector. This is then fed as input to the proposed Bi-Attention LSTM model for classification.

C. Classification - Bi-Attention LSTM model

Bi-Attention LSTM is an extended version of the traditional



(a)



(b)

Fig. 3. Architecture of (a) LSTM (b) Bi-Attention LSTM

The activation functions of LSTM are calculated for Input sensor data I_t , output predicted sensor data O_t , forget sensor data F_t , candidate vector C'_t , cell state C_t and hidden state h_t , using the following formulae,

$$I_t = \sigma(W_i[X(t), h_{t-1}] + b_i) \quad (6)$$

$$F_t = \sigma(W_f[h_{t-1}, X(t)] + b_f) \quad (7)$$

$$O_t = \sigma(W_o[h_{t-1}, X(t)] + b_o) \quad (8)$$

$$C'_t = \sigma(W_c[h_{t-1}, X(t)] + b_c) \quad (9)$$

$$C_t = F_t * C_{t-1} + I_t * C'_t \quad (10)$$

LSTMs. It is capable of improvising the working ability of the trained model on the sequence grouping issues. These Bi-Attention LSTM models train two LSTMs rather than one on the obtained input sequence. All of the input sequence time-steps are available here. The main idea underlying this proposed methodology is simple. It duplicates the primary, recurrent layer of the network. The sequence is then given as an input to the corresponding primary layer and then giving the input sequence flipped copy to the subsequent duplicated layer. Fig. 3, shown below, illustrates the LSTM and Bi-Attention LSTM architecture.

Regarding Fig. 3, it is observed that Bi-Attention LSTM is an improvisation of LSTM and comprises an input, output, and a hidden layer. Bi-Attention LSTMs, as previously said, train two LSTMs rather than one on an input sequence, as seen in Fig. 3.

$$h_t = O_t * \tanh(C_f) \quad (11)$$

Where $X(t)$ is the vector of input data, h_{t-1} is the prior state vector, W is the weight, and b is the bias for every sensor data. This proposed system aids in overcoming the shortcomings of the existing methods, all of the accessible input data in the past and in the future of a particular time-step can be used to train the Bi-Attention LSTM. Furthermore, a split of state neurons in a conventional RNN is responsible for a portion of the backward state that signals the negative time direction and is liable for forwarding states representing the positive time direction. Thus, this proposed system classifies the multi-task activities of the humans from the video as normal or abnormal.

IV. RESULTS AND DISCUSSION

CNN with Bi- Attention LSTM is used to identify human beings' normal and abnormal activities automatically. The performance of the proposed method is validated by testing the ability to detect abnormalities in a UMN dataset. Input is taken as a video format, and then the videos are converted into video frames and removed the unwanted things. The pre-processed images have proceeded for feature extraction. SpatialTemporalNet is used to extract the features. Spatial-Temporal Net contains peculiar pooling and convolutional layers, so the foreground images are very effectively extracted. These images are fed in the Bi-Attention LSTM classifier to predict the correct class.

Dataset description

The proposed features and applications are evaluated using the UMN Database (UMN) of the University of Minnesota in the United States. The UMN database contains unusual and normally crowded videos. The UMN database has a resolution of 240 * 320 pixels. The database, which is used extensively in detecting inconsistent behavior, accurately calculates many aspects such as the size of the population in

different situations, variable lighting, and the uncertainty and inconsistency of crowd movements. It features three peculiar scenarios of escape events in different outdoor and indoor scenes. With each of these situations, the collection of individuals usually takes place in one area. Suddenly, all the individuals (escape) run away. Here, escape is taken into account as a paradox [32].

A. Experimental results

The implementation of the proposed system classifies the multi-task activities of humans from videos as normal or abnormal through feature extraction and classification. The classified results are shown in table 1 and table 2. As per these results, humans perform multi-task activities, which can be perceived from the videos. But, classifying it as abnormal and normal is important to avoid any unpleasant events in public places. Table 1 shows the normal events, while table 2 shows the abnormal events. The figures are also presented in grayscale.

TABLE 1. RGB TO GREYSCALE CONVERSION FOR NORMAL EVENTS

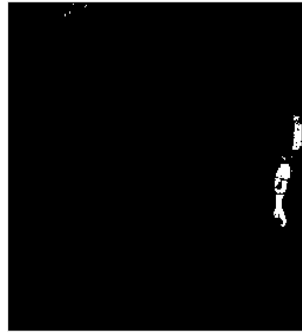


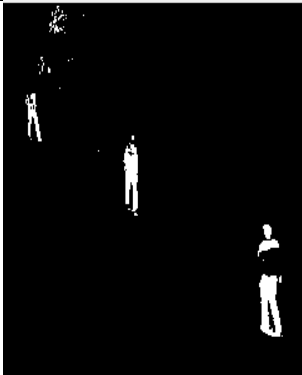





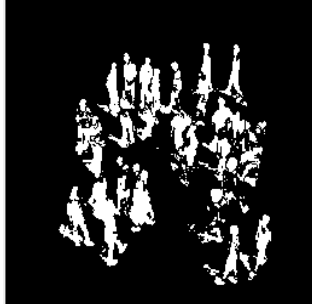

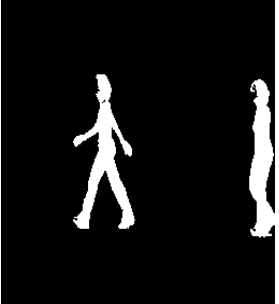
Original image	Grayscale image	Original image	Grayscale image
			
			

TABLE 2: RGB TO GREYSCALE CONVERSION FOR ABNORMAL EVENTS

Original image	Grayscale image	Original image	Grayscale image
			
			

RGB to greyscale conversion for normal and abnormal events is provided in table 1 and table 2, respectively. The proposed method perfectly converts the RGB images into Greyscale images. These images were given for classification for class prediction. The proposed Bi-Attention LSTM classifier analyzes the input images and predicts an accurate class. The performance of the proposed CNN integrated two-focus LSTM system has been analyzed and verified in the following section.

A. Performance Analysis

UMN dataset does not include pixel-level ground reality, and discrepancies are staged. This dataset contains both normal and abnormal activities. Proposed Bi-Attention LSTM is used to find the actions perfectly. The performance of the proposed system is analyzed in terms of accuracy, sensitivity, specificity, error, precision, F1-score, FPR, Kappa, and MCC. The proposed Bi-Attention LSTM is compared with the existing VGG16, ResNet, Alexnet, and Googlenet with respect to the nine metrics.

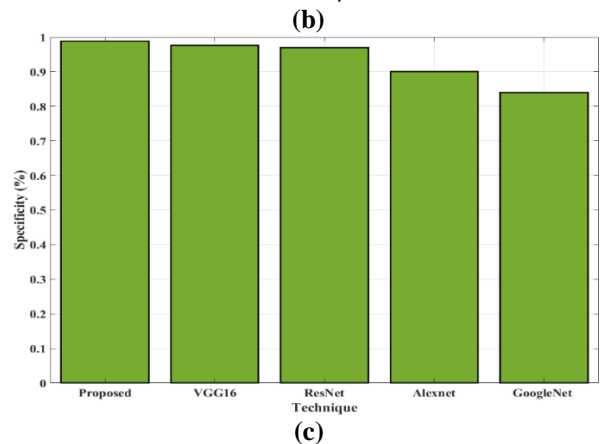
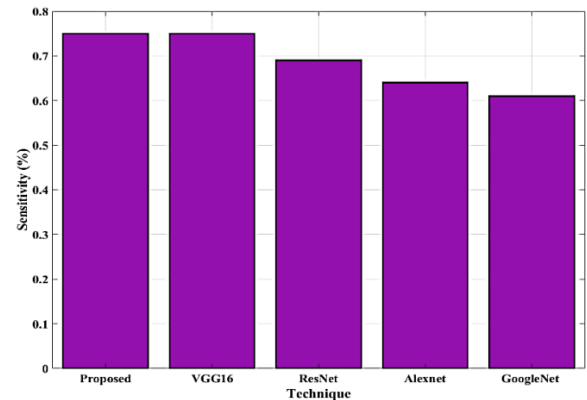
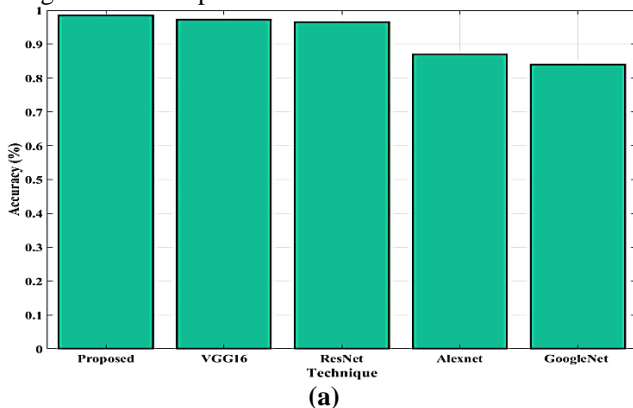


Fig. 4. Comparison of Proposed and existing methods (a) Accuracy (b) Sensitivity (c) Specificity

Initially, the Accuracy of proposed and existing methods are compared. The accuracy of the proposed system is found to be 98.44%, but the existing methods of VGG16 are found to

be 97.28%, ResNet is found to be 96.50%, Alexnet is found to be 87%, and Googlenet is found to be 84%. Likewise, the sensitivity of the proposed and existing methods are analyzed. Sensitivity is important for perceiving network traffic because it accurately measures a positive data set. The proposed method has 75% sensitivity, but the existing methods of VGG16 has 75%, ResNet has 69%, Alexnet has 64%, and Googlenet has 61%. Similarly, specificity is also analyzed. Specificity is one of the statistical methods for correctly recognizing a negative data set. In comparing existing methods, the proposed method has 98.81% specificity, but VGG16 has 97.63%, ResNet has 96.97%, Alexnet has 90%, and Googlenet has 84%. Fig. 4 illustrates the comparison of proposed and existing methods.

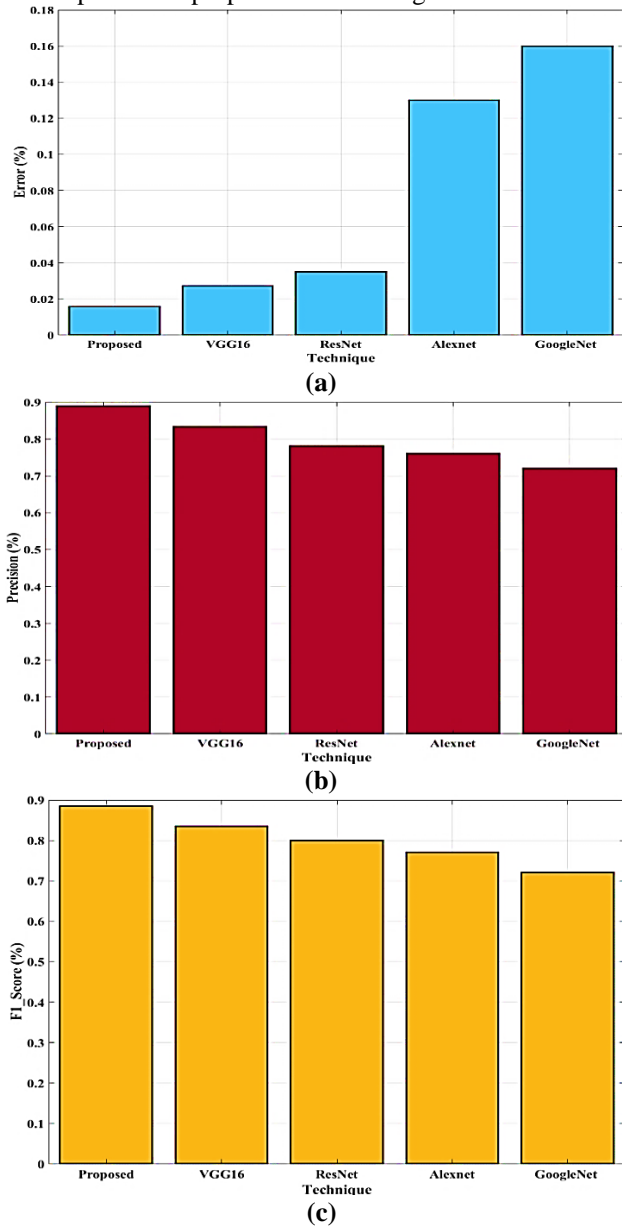


Fig. 5. Comparison of the proposed and existing methods (a) Error (b) Precision (c) F1_Score

Furthermore, error, precision, and F1_score are analyzed that as shown in Fig. 5. The discrepancy between the calculated and actual values is referred to as error in statistical analysis. The proposed method has a low error value of 1.56%, but the existing methods of VGG16 have 2.72% error, ResNet has 3.50% error, Alexnet has 13% error, and Googlenet has 16% error. Similarly, the precision and F1_score are also analyzed and verified. The precision of the proposed method is found to be 89% and 88.50% of the F1_score value.

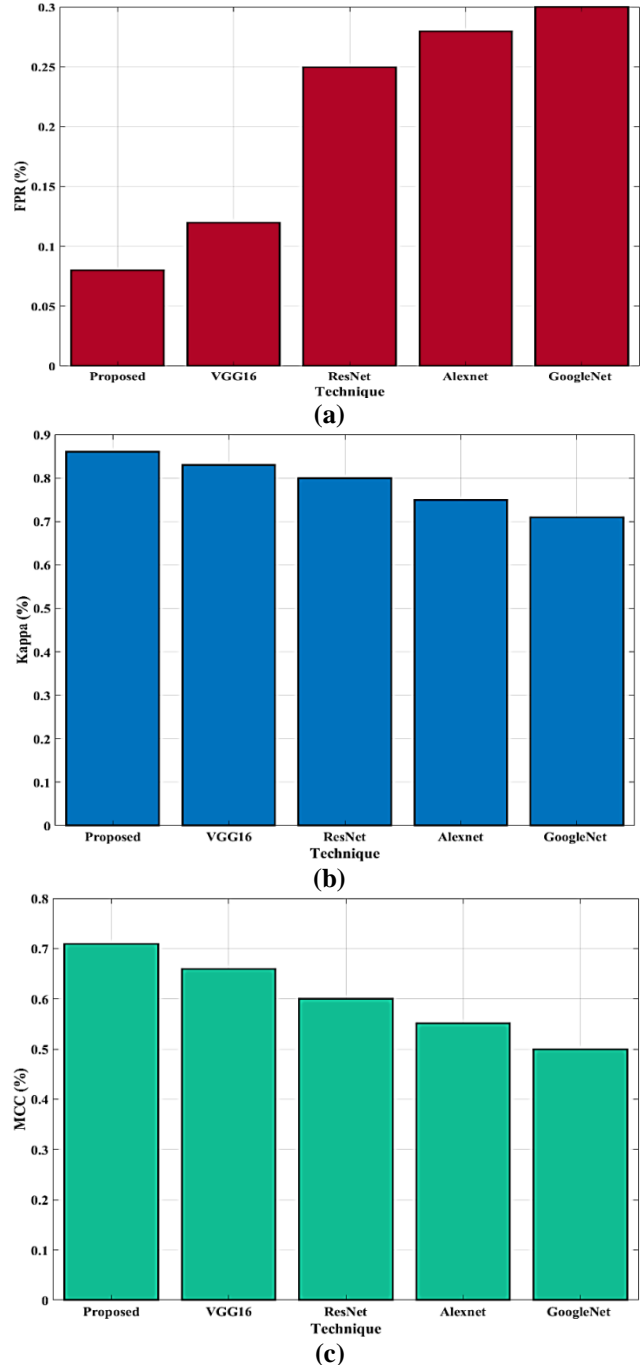


Fig. 6. Comparison of proposed and existing methods (a) FPR (b) Kappa (c) MCC

Moreover, False Positive Rate (FPR), Kappa, and Mathews correlation coefficients were also analyzed and verified, illustrated in Fig. 6. The relationship between the number of negative events incorrectly classified as positive and the total number of true negative events is used to compute the false positive rate. The proposed method has a low false-positive rate of 8%, but VGG16 has the value of FPR as 12%, ResNet has 25% of FPR, Alexnet has 28% FPR, and Googlenet has 30% FPR. Similarly, the kappa value of proposed and existing methods are also analyzed. The statistical analysis of Kappa is described as comparing the observed values of a training dataset to the expected value. In the proposed method, the value of kappa is 86%, 83% in VGG16, 80% in ResNet, 75% in Alexnet, and 71% in Googlenet. Mathews' correlation coefficient is then analyzed, and, in machine learning, the Mathews correlation coefficient (MCC) is used to estimate the validity of two binary classifications. MCC value of the proposed method is 71%, but the existing VGG16, ResNet, Alexnet, and Googlenet MCC values are 66%, 60%, 55.2%, and 50%, respectively. Table 3 shows the overall performance of the proposed and existing methods.

TABLE 3. OVERALL PERFORMANCE ANALYSIS OF THE PROPOSED AND EXISTING SYSTEM

Performance metrics	Proposed Bi-Attention LSTM	VGG 16	Res net	Alex net	Go gle net
Accuracy	0.9844	0.972	0.9	0.87	0.8

TABLE 4. COMPARISON OF TRADITIONAL TECHNIQUES [31] ON THE UMN DATASET

Methods	SF	H-MDT CRF	AVID	GANs	Deep-Cascade	Anomaly Net	Proposed System
EER	12.60%	3.70%	2.60%	-	2.50%	2.60%	1.56%
AUC	94.90%	99.50%	99.60%	99.00%	99.60%	99.60%	99.88%

From table 4, the AUC rate of the presented model is more than the existing methods. On the other hand, the EER that is the Equal Error Rate is minimum or the proposed system than the other traditional methods. This shows the efficacy of the proposed method as it explores high AUC and minimum error rates on the UMN dataset. Similarly, an internal comparison is made for the proposed system for various epochs. This is tabulated for both normal and abnormal events and presented in table 5 and table 6. As per table 5, when the epoch is 50, the Tanh value is 94.75. Similarly, when the epoch is 150, it explores a value of 94.16. This value increases as the epochs increase. LeakyReLU and ReLU also show the same outcomes. That is, this rate increases as the epochs increase. On the other hand, the abnormal event also shows results in a way that all the activation function increases with respect to increase in epochs, as shown in table 6.

		8	650		4
Sensitivity	0.75	0.75	0.69	0.64	0.61
Specificity	0.9881	0.9763	0.9697	0.9	0.84
Error	0.0156	0.0272	0.0350	0.13	0.16
Precision	0.890	0.834	0.7813	0.76	0.72
F1_score	0.8850	0.835	0.80	0.77	0.72
FPR	0.08	0.12	0.25	0.28	0.3
Kappa	0.86	0.83	0.8	0.75	0.71
MCC	0.71	0.66	0.6	0.552	0.5

The traditional study [31] considered ResNet, VGG16, Alexnet, and Googlenet to detect video anomalies. The performance of these techniques is found to be minimum in comparison to the proposed system. Thus, the proposed system is more effective than traditional methods for classifying human activities as abnormal or normal. The proposed Bi-Attention LSTM is more effective than the other four methods due to the outstanding accomplishment of the proposed system in terms of accuracy, sensitivity, specificity, error, precision, F1_score, FPR kappa, and MCC.

TABLE 5. NORMAL EVENT

Normal Event				
Epochs	50	150	250	350
Activation Function				
Tanh	94.75	94.16	94.98	95.15
Leaky ReLU	95.21	95.74	95.32	96.37
ReLU	95.34	95.68	95.71	95.55

TABLE 6. ABNORMAL EVENT

Abnormal Event				
Epochs	50	150	250	350
Activation Function				
Tanh	92.51	92.45	92.68	92.49
Leaky ReLU	93.11	93.64	93.85	94.05
ReLU	93.24	93.36	93.33	94.08

Hence, it is clear that the presented model is effective than the traditional method with respect to the six significant performance metrics, which are explored from the comparison results. Similarly, the internal comparison (table 5 and 6) also shows effective outcomes as epochs increase. This proves the efficacy and accuracy of the proposed system for classifying the human activities from the videos as normal and abnormal.

V. CONCLUSION

Important to categorize many of the tasks of humans from videos to avoid suspicious actions. Manual detection techniques do not accurately detect abnormal functions because abnormal behaviors rarely occur, so automated detection techniques have been developed to identify abnormal functions at that time. The proposed work has introduced a convolutional neural network (CNN) and two-focus long-term memory (two-focus LSTM) to detect normal and abnormal human functions automatically. Here, spatial and temporal features are extracted and classified by the proposed Bi-Attention LSTM to predict the normal and abnormal events. Videos are taken as input and then converted to frames. Following that, the frame is sent into a CNN-based Bi-Attention LSTM to predict normal and abnormal actions. The proposed methodology is analyzed to find the extent to which the proposed system is better than the existing methods. For this purpose, nine performance metrics have been considered, and the analysis is carried out by comparing the proposed system with traditional ResNet and VGG 16, Alexnet, and Googlenet. Comparison is also undertaken by considering the UMN dataset. The proposed method attained 98.44% accuracy with 8% FPR. Internal comparison is also performed by taking the epochs into account. The performance analysis explored efficient outcomes of the proposed system for classifying the multi-task activities of humans from videos as normal and abnormal. Extensive testing demonstrates that the proposed system works best in image restoration and irregular event detection on CCTV. In future work, an improved Residual network (ResNet) will be added to the existing model configuration to expand the number of network layers.

REFERENCES

- [1] V. Tsakanikas and T. Dagiuklas, Video surveillance systems-current status and future trends, *Computers & Electrical Engineering*, 70 (2018) 736-753.
- [2] S. Ramasamy Ramamurthy and N. Roy, Recent trends in machine learning for human activity recognition—A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8 (2018) e1254..
- [3] W. Xu, Z. Miao, X.-P. Zhang, and Y. Tian, A hierarchical Spatio-temporal model for human activity recognition, *IEEE Transactions on Multimedia*, 19 (2017) 1494-1509.
- [4] A. Ullah, K. Muhammad, W. Ding, V. Palade, I. U. Haq, and S. W. Baik, Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications, *Applied Soft Computing*, 103 (2021) 107102.
- [5] V. M. Vishnu, M. Rajalakshmi, and R. Nedunchezian, Intelligent traffic video surveillance and accident detection system with dynamic traffic signal control, *Cluster Computing*, 21 (2018) 135-147.
- [6] Y. Li, R. Xia, Q. Huang, W. Xie, and X. Li, Survey of Spatio-temporal interest point detection algorithms in the video, *IEEE Access*, 5 (2017) 10323-10331.
- [7] S. Chaudhary, M. A. Khan, and C. Bhatnagar, Multiple anomalous activity detection in videos, *Procedia Computer Science*, 125 (2018) 336-345.
- [8] D. Singh and C. K. Mohan, Graph formulation of video activities for abnormal activity recognition, *Pattern Recognition*, 65 (2017) 265-272.
- [9] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, Activity recognition using temporal optical flow convolutional features and multilayer LSTM, *IEEE Transactions on Industrial Electronics*, 66 (2018) 9692-9702.
- [10] F. Zhou, L. Wang, Z. Li, W. Zuo, and H. Tan, Unsupervised learning approach for abnormal event detection in surveillance video by hybrid autoencoder, *Neural Processing Letters*, 52 (2020) 961-975.
- [11] M. Al-Nawashi, O. M. Al-Hazaimeh, and M. Sarace, A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments, *Neural Computing and Applications*, 28 (2017) 565-572..
- [12] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, Suspicious human activity recognition: a review, *Artificial Intelligence Review*, 50 (2018) 283-339..
- [13] A. B. Mabrouk and E. Zagrouba, Abnormal behavior recognition for intelligent video surveillance systems: A review, *Expert Systems with Applications*, 91 (2018) 480-49.
- [14] C. Dhiman and D. K. Vishwakarma, A review of state-of-the-art techniques for abnormal human activity recognition, *Engineering Applications of Artificial Intelligence*, 77 (2019) 21-4.
- [15] H. Hattori, N. Lee, V. N. Boddeti, F. Beainy, K. M. Kitani, and T. Kanade, Synthesizing a scene-specific pedestrian detector and pose estimator for static video surveillance, *International Journal of Computer Vision*, 126 (2018) 1027-1044.
- [16] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, Spatiotemporal anomaly detection using deep learning for real-time video surveillance, *IEEE Transactions on Industrial Informatics*, 16 (2019) 393-402.
- [17] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, Robust human activity recognition from depth video using spatiotemporal multi-fused features, *Pattern Recognition*, 61 (2017) 295-308.
- [18] K. Pawar and V. Attar, Deep learning approaches for video-based anomalous activity detection, *World Wide Web*, 22 (2019) 571-601.
- [19] W. Huang, H. Ding, and G. Chen, A novel deep multi-channel residual networks-based metric learning method for moving human localization in video surveillance, *Signal Processing*, 142 (2018) 104-113.
- [20] W. Bouachir, R. Gouiaa, B. Li, and R. Noumeir, Intelligent video surveillance for real-time detection of suicide attempts, *Pattern Recognition Letters*, 110 (2018) 1-7.
- [21] H. Yao, A. Cavallaro, T. Bouwmans, and Z. Zhang, Guest editorial introduction to the special issue on group and crowd behavior analysis for intelligent multi-camera video surveillance, *IEEE Transactions on Circuits and Systems for Video Technology*, 27 (2017) 405-408.
- [22] T. Wang, M. Qiao, Y. Deng, Y. Zhou, H. Wang, Q. Lyu, et al., Abnormal event detection based on analysis of movement information of video sequence, *Optik*, 152 (2018) 50-60.
- [23] A. Jordao, L. A. B. Torres, and W. R. Schwartz, Novel approaches to human activity recognition based on accelerometer data, *Signal, Image and Video Processing*, 12 (2018) 1387-1394.
- [24] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, Deep learning models for real-time human activity recognition with smartphones, *Mobile Networks and Applications*, 25 (2020) 743-755.
- [25] X. Zhang, Q. Yu, and H. Yu, Physics inspired methods for crowd video surveillance and analysis: a survey, *IEEE Access*, 6 (2018) 66816-66830.
- [26] F. Najar, S. Bourouis, N. Bouguila, and S. Belghith, Unsupervised learning of finite full covariance multivariate generalized Gaussian mixture models for human activity recognition, *Multimedia Tools and Applications*, 78 (2019) 18669-18691.

- [27] T. Singh and D. K. Vishwakarma, Video benchmarks of human action datasets: a review, *Artificial Intelligence Review*, 52 (2019) 1107-1154.
- [28] K.-E. Ko and K.-B. Sim, Deep convolutional framework for abnormal behavior detection in a smart surveillance system, *Engineering Applications of Artificial Intelligence*, 67 (2018) 226-234 .
- [29] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie, and J. Zhou, Tse-cnn: A two-stage end-to-end cnn for human activity recognition, *IEEE journal of biomedical and health informatics*, 24 (2019) 292-299.
- [30] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan, et al., Human action recognition using a fusion of multiview and deep features: an application to video surveillance, *Multimedia tools and applications*, (2020) 1-27.
- [31] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, AnomalyNet: An anomaly detection network for video surveillance, *IEEE Transactions on Information Forensics and Security*, 14 (2019) 2537-2550.
- [32] Al-Dhamari, A., Sudirman, R. and Mahmood, N.H., Transfer deep learning along with binary support vector machine for abnormal behavior detection. *IEEE Access*, 8 (2020) 61085-61095.