Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PERFORMANCE ANALYSIS OF LSTM BASED DEEP LEARNING MODELS FOR ABNORMAL ACTION PREDICTION IN SURVEILLANCE VIDEOS

MRS. MANJU D[1], DR. SEETHA M[2], DR. SAMMULAL P [3]

[1]Assistant Professor, Dept.of CSE, GNIITS, Hyderabad, India
[2]Professor & HOD, Dept. of CSE GNIITS, Hyderabad, India
[3]Professor & Dept.of CSE, JNTUH CEJ, Hyderabad, India

E-mal: [1]s.r.manju@gnits.ac.in, [2] maddala.seetha@gnits.ac.in, [3] sam@jntuh.ac.in

ABSTRACT

Video surveillance is increasingly being adopted for ensuring safety and security both in public and private places. Automated prediction of abnormal events like theft, robbery, murder etc from continuous observation of surveillance videos is a multidisciplinary study involving computer vision, deep learning and artificial intelligence. Deep learning-based video analysis and categorization is the most researched topic. Many deep learning models based on Long Short Term Memory are proposed for automated prediction of abnormal events. There are two contributions in this paper; the first contribution focuses on five models - Resnet, VGG16, VGG19, 3DCNN and Inception V3. The second contribution has proposed an approach called Recurrent-Residual-Inception V3 (RRIV3). Advantage of RRIV3 is performance will not get effected more by removal of any residual block. This work does a performance analysis of six LSTM based deep learning models for abnormal event prediction from surveillance videos before and after performing preprocessing. Deep learning models are combined with LSTM for the prediction of abnormal events from past observation of events in the video stream. These six models are executed against different benchmarked abnormal event detection datasets one among them is UCF-Crime dataset and efficiency is compared in terms of accuracy, precision, recall and execution time. It is observed that Recurrent-Residual-Inception V3 with LSTM performs better than other models with training accuracy of 90% and test accuracy of 85% compared to other models. The execution time is 20 milliseconds compared to other models.

Keywords: *3DCNN, Inception V3, VGG16, VGG19, Resnet, Recurrent-Residual-Inception V3*

1. INTRODUCTION

Video surveillance systems are deployed in many places like roads, stations, airports, malls etc. for public safety, however detecting abnormal activities and taking proactive actions can provide better security to individuals. For this, people and their interactions must be constantly monitored for a longer duration and any abnormal activity must be predicted. It is difficult for trained personnel to reliably monitor videos for a longer duration and predict abnormal events. With the need to automate this activity with high accuracy, many autonomous abnormal activity detection systems are proposed. The goal of any autonomous anomaly recognition system is to detect/predict any offensive or disruptive activities in the surveillance video in real-time. The conventional systems extract various features of appearance, dynamic relationships and interactions between the entities in the video and classify them to detect any abnormal activity. The accuracy is limited in this approach due to the insufficiency of handcrafted features to detect abnormal activity. As abnormality is context-dependent, the identification of features that represent the activity in the relevant context is challenging. Recently deep learning algorithms are being used for many computer vision problems. Deep learning algorithms learn features automatically and provide better accuracy. Deep learning uses discriminative feature representations of both appearance and motion patterns to model the event patterns.

In this work, Deep learning models are used for the classification of objects within the frame, a comparative analysis of six deep learning LSTM based models for abnormal event prediction is presented. With the capability of learning long term dependencies and the ability to extrapolate temporarily sequential data, long short term memory (LSTM) are best suited for abnormal event prediction. LSTM combined with deep learning event classification can provide better accuracy of abnormal event prediction. This work explores six different deep learning models of Resnet, VGG16, VGG19, 3DCNN (3D deep convolutional neural network), Inception V3 and Recurrent-Residual-Inception V3 in combination with LSTM for abnormal event prediction. The performance of these six models is compared against benchmarking datasets in terms of Accuracy, Precision, Recall and execution time.

Existing research methodologies considered abnormal action detection where different authors proposed differently like a novel spatio-temporal U-Net for frame prediction using normal events and abnormality detection using prediction error. Used U-Nets to represent spatial data and ConvLstm to represent motion data. In addition, authors propose a new regular score function, consisting of a prediction error for not only the current frame but also future frames, to further improve the accuracy of anomaly detection [1]. An approach called S^2–VAE is used, which is a combination of two neuralnetworks [2]. Spatio-temporal generator and discriminator along with Bi-directional ConvLSTM is used to determine whether the given video sequences are normal or abnormal [3]. Super orientation optical flow, are designed to easily replicate motion info [4]. Used a similar approach wherein they use optical flow and temporal frames, they quantize them into 7 bits and then map these bits into histograms, histograms of normal frames would conform to a similar distribution, whereas an abnormal one would have a very different representation[5]. MONAD framework is used, to provide location, motion ,bounding and appearance for that a threshold is set to check false alarm rate[6]. CNN is used to extract spatial features and these are given as input to Residual LSTM to recognize anomalous events in video surveillance[7]. Used DNN for future frame prediction and that frame is compared with the testing video frame for similarity[8]. Presented a framework that recognizes abnormal human behavior by combining deep convolutions with standard RGB images. Using the unified structure would allow detection speed to be improved while ensuring accuracy. Human-oriented deep convolutional framework is composed of (i) a detection and discrimination module that provides a means for distinguishing different entities, using a different algorithm from existing methods, and (ii) a module to classify abnormal postures by using spatial features, and (iii) a module for detecting abnormal behavior using long short-term memory (LSTM)[9]. Two stream CNN architecture take advantage of both anomalous and nonanomalus films[10]. But still improvisation is needed to give the correct anomaly detection, few adopted pixel-level information-based anomaly detection, which is a complex feature extraction process that requires a huge amount of data and a lower speed of execution. Many factors affect accurate abnormality detection, such as variations in luminance and shadows. If the luminance and shadows problem is not corrected, then the information extraction is very difficult.

2. DEEP LEARNING BASED LSTM MODELS

The six deep learning based LSTM models used for abnormal activity prediction is detailed in this section. LSTM is combined with Resnet, VGG16, VGG19, 3DCNN, Inception V3 and Recurrent-Residual-Inception V3 along with pre-processing step.

The work involves

Step1: Pre-processing
- Frame conversion pre-processing
- Background removal
- Motion estimation & Object tracking

Step2: Classification using deep learning models specified for human action and event categorization.

Step3: Prediction using LSTM

Step4: To compare and analyse metrics like accuracy, precision, recall, execution time with and without pre-processing.

Initially, the input video is converted into frames. After that, the pre-processing step will be carried out.

In pre-processing, Histogram Equalization (HE) technique is compared with technique called

Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Histogram Partition based Gamma correction. The gamma correction can be defined as.

$$v_{out} = ay(i,j)^r \quad (1)$$

Where, the non-negative real input value $y(i,j)$ is raised to the power r and multiplied with the constant a to obtain the output value v_{out}.

The existing histogram technique returns the over enhanced problem due to range-based pixel selection. In order to solve this issue in pre-processing, the intensity level will be adjusted by gamma correction.

Next, the background will be removed by using the Modified Livewire Segmentation (MLWS) algorithm. In this step, the exiting Livewire segmentation algorithm will be done by using Sobel Filter and the gradient magnitude of the edges degrades in this filter which leads to inaccurate results. In order to solve this problem in existing Livewire Segmentation, Bilateral Filter is replaced with Sobel Filter for better performance.

In order to improve the Livewire algorithm, the bilateral filters are used as edge detectors. The choice of bilateral filter is due to its edge preserving and smoothness in noise reduction. Each pixel is replaced with its weighted average of its neighbourhood pixels intensity. The pixel value is calculated as

$$\lambda(i,j,o,l) = Exp\left(\frac{(i-o)^2}{2\sigma_{Di}^2} - \frac{\left\|\Im(i,j)-\Im(o,l)^2\right\|}{2\sigma_{Ri}^2}\right) \quad (2)$$

Where, σ_{Di} and σ_{Ri} defines the smoothing parameters, $\Im(i,j)$ and $\Im(o,l)$ defines the intensity of pixels respectively.

After that, for Motion estimation used Modified Adaptive Root pattern search (MARPS). In this, distance standard deviation is used instead of mean absolute difference. Distance standard deviation (DSD) is not affected by extreme values.

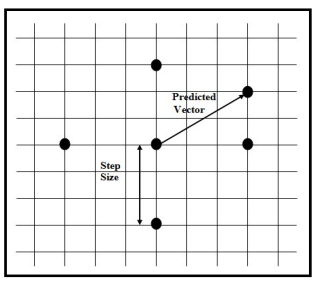

Figure 1: Adaptive Root Pattern Search Method

The steps in DSD values can be calculated as

$$DSD = \sqrt{(C_{F1} - C_F ban)^2 + (R_F - C_F ban)^2} \quad (3)$$

If DSD < threshold value means motion-vector (Mv) remains unchanged, otherwise,

$$Ps = Max\{|Mv\,predicted(X), Mv\,predicted(Y)|\} \quad (4)$$

Mean square error is given as

$$MSE = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(C_{F_{ij}} - R_{F_{ij}})^2 \quad (5)$$

Here, C_F indicates current frame, and R_F indicates reference frame. Center point is iteratively replaced with the minimum error point found in earlier step, until a minimum error criteria is reached.

After that, the object will be detected by using YOLOV4. The steps of YOLOV4 algorithm are: Initially, the frames are divided into multiple grid cells, and then the anchor boxes are defined. Anchor boxes are predefined orientations of a 2D box that might expect to find an object. Furthermore, a bounding box prediction is also defined. The bounding box for the object is predicted and its coordinates are provided as output by YOLOV4. The bounding box is defined in terms of starting x , y coordinates, width and height as $(x_{ctr}, y_{ctr}, wd, ht\}$ with prediction expression given as

$$B_x = \sigma(T_x) + C_x \quad (6)$$

$$B_y = \sigma(T_y) = C_y \quad (7)$$

$$B_{wd} = P_{wd}.e^{T_{wd}} \quad (8)$$

$$B_{ht} = P_{ht}.e^{T_{ht}} \quad (9)$$

Where, P_{wd} and P_{ht} is the width and height of the anchor box. (C_x, C_y) is the coordinate of the top left corner of the image, $(B_x, B_y, B_{wd}, B_{ht})$ denotes the dimensions of our predicted bounding box.

Once the objects get tracked, from which the most important and required features are extracted out. In this work, two kinds of features are extracted. One is low features and another is high

features. The feature extraction process is mathematically defined as.

$$\lambda_{EX}(\varphi_d) = (\lambda_i^{Low}, \lambda_i^{High}) \quad (10)$$

Where, $\lambda_{EX}(\varphi_d)$ denotes the feature extraction of detected object, λ_i^{Low} denotes low features, and λ_i^{High} denotes high features.

The low features resemble the features like Histogram of optical flow λ_{HO} and Histogram of gradient λ_{HG}, and that function is expressed as.

$$\lambda_i^{Low} = \{\lambda_{HO}, \lambda_{HG}\} \quad (11)$$

The high features resemble the features such as Shape λ_s, Texture λ_T, Surf λ_{Sf}, SIFT λ_{SI}, Boundary box λ_{Bb}, and human skeleton features λ_{Hs}.

$$\lambda_i^{High} = \{\lambda_S, \lambda_T, \lambda_{Sf}, \lambda_{SI}, \lambda_{Bb}, \lambda_{Hs}\} \quad (12)$$

Next, for the classification used models+ LSTM (like, VGG16, VGG19, 3DCNN, InceptionV3, Recurrent-Residual-Inception V3 +LSTM). Using the metrics like accuracy, precision, Recall. The results are compared among the all deep learning classification models like VGG16, VGG19, 3DCNN, ResNet50, InceptionV3 and Recurrent-Residual-Inception V3.Created an extra layer called custom activation for Kernel function for improving the classification accuracy.

The specified models are combined with LSTM to predict the anomaly or non-anomaly person before the action takes place.

LSTM extends recurrent neural network to solve the vanishing gradient problem. Memory units in LSTM store the learnt knowledge.

LSTM has three gates of input(i), forget(f) and output(o) to forget already learnt and update newly learnt knowledge.

The update and output functions are defined as below

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + b_i) \quad (13)$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + b_f) \quad (14)$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + b_o \quad (15)$$

$$g_t = \sigma(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \quad (16)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad h_t = o_t \odot c_t \quad (17)$$

$$\sigma(x) = (1 + e^{-x})^{-1} \quad (18)$$

In the above equations, the non-linear sigmoid function is given as σ(x). The matrix encoding the gate's parameters is given as W. The gates are controlled adaptively to solve the problem of vanishing gradients.

The overall structure of the comparison models is given in Figure 2.

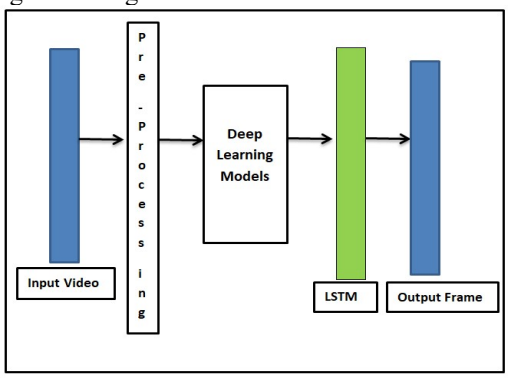

Figure 2: Deep Learning LSTM Model

Deep learning models extract features and provide to LSTM for prediction of abnormality. Six deep learning models of Resnet, VGG16, VGG19, 3DCNN, Inception V3 and Recurrent-Residual-Inception V3 are used for extracting features.

2.1 Resnet with LSTM

Resnet or Residual Network was proposed by Microsoft researchers in 2015 as a solution to the problem of vanishing/exploding gradient with the increase in number of layers in deep convolutional neural network. Skip connection strategy is adopted in the Resnet network to enable feature learning ability. Due to this the depth of the network is expanded leading to improved learning ability. Gradient vanishing problem is solved effectively in Resnet due to passage of useful information to next layer with skip connection model.

Resnet with LSTM abnormality prediction model uses Resnet50 for spatial feature extraction and LSTM for prediction using temporal feature extraction. The Resnet-50 is trained using transfer learning strategy to boost the training

performance. The advantage of transfer learning it that, it is able to reduce the number of parameters to be trained and achieved convergence faster.

2.2 VGG16 with LSTM

VGG16 is a simple deep convolutional neural network. It is formed by stacking deep CNN followed by two fully connected layers. The fully connected layers have 4096 neurons on each of them. In VGG16 with LSTM model, VGG16 network prepares the input feature vector for LSTM, which predicts abnormality.

Input frames from the video are sized to 224*224. The features are extracted using VGG16 and provided to LSTM in a time sequence, to provide abnormal or normal class as output.

In VGG16 model the dimensions of the extracted features are (7, 70,512). The feature vector is sliced into a set of 10 samples with each of size {7, 7*512}. This feature vector is given as input to LSTM with 10 time steps. The output of the last cell in LSTM is a binary output which gives as 1 for abnormal and 0 for normal.

2.3 VGG19 with LSTM

VGG19 is similar to VGG16 but it is deeper than VGG16. Due to this depth, VGG19 is able to extract more low-level features from a frame than VGG16. VGG19 has 16 layers for convolution, 3 layers fully connected, 5 Max pool layer, and 1 softmax layer. Feature extraction part in this architecture if from input layer to the last max-pooling layer (7 x 7 x 512). The feature extracted from VGG19 model is passed to LSTM to learn the long-term dependencies between the video frames and predict abnormal activity. The pipelining of VGG19 with LSTM is same as that used for VGG16.

2.4 3DCNN with LSTM

3DCNN has total of 20 layers with 12 for convolution, 5 for pooling and 1 for fully connected. The remaining 2 is for LSTM and output. At each convolutional layer in 3DCNN, a 2D CNN, pooling and drop layers are paired. A drop ratio of 0.25 is set in this network. 2D CNN uses 3*3 kernels with ReLU activation function. LSTM is used for extracting temporal correlation. After temporal analysis, class of frame (normal/abnormal) is predicted at the output layer.

2.5 Inception V3

The Inception V1 also called GoogLeNet produced lowest error for ImageNet classification dataset. But use of 5*5 convolutions cause decrease of input dimension and resulted in reduced accuracy. To improve it Inception V2 architecture replaced 5*5 convolution with two 3*3 convolution. In addition to increased accuracy, this change also reduced the computation time by 2.78 times compared to Inception V1. For making to work as inception V3, the changes had been made to Inception V1. The first change is Factorization in to smaller convolutions named as Inception Block A then added blocks named Inception Block B and C which performs asymmetric factorization. It's a 48 layer architecture.

The feature extracted from Inception V3 model is passed to LSTM to learn the long term dependencies between the video frames and predict abnormal activity. The pipelining of Inception V3 with LSTM is same as that used for VGG16.

2.6 Recurrent-Residual-Inception V3

Recurrent-Residual-Inception V3 is the proposed method that have been central to the largest advances in image recognition performance in recent years because the model is increased in terms of depth and width, which results in very good performance at a relatively low computational cost.

The extracted features λ from the frames are given to the classifier via input layer, and that function is mathematically formulated as.

$$\lambda_{ip} = \begin{bmatrix} \lambda_{1\times1} & \lambda_{1\times2} & \lambda_{1\times3} \\ \lambda_{2\times1} & \lambda_{2\times2} & \lambda_{2\times3} \\ \lambda_{3\times1} & \lambda_{3\times2} & \lambda_{3\times3} \end{bmatrix} \quad (19)$$

The problem of the training process and the vanishing gradient problem can be resolved by using the residual network. In this network, the skip connection technique is used. The skip connection skips few layers and connects directly to the output.

Without using this skip connection, the input λ gets multiplied by the weights of the layer followed by summing up with bias term. Then, the term goes through the activation function, $F()$ and the output is obtained as $\delta(\lambda)$.

$$\delta(\lambda) = F(\varpi\lambda + \beta_i) \quad (20)$$

$$or \ \delta(\lambda) = F(\lambda) \quad (21)$$

After the skip connection, the output is transformed to

$$\delta(\lambda) = F(\lambda) + \lambda \quad (22)$$

When the dimensions of the inputs vary from that of the output, which can happen with convolutional and pooling layers. Such that the dimensions of $F(\lambda)$ are different from λ, in this case, the following criteria is used.

- Skip connection: means adding extra zero entries for increasing the dimension.
- Projection method: matches the dimension by adding 1x1convolutional layers to the input. In such a case, the output is:

$$\delta(\lambda) = F(\lambda) + \varpi 1.\lambda \quad (23)$$

Recurrent is defined as $O_i=R_i(Z_{i-1})+Z_{i-1}$ where R_i is residual block, Z_{i-1} is operation on every input sequence, which improves accuracy as the input has to follow the sequential path at time t. RRIV3 combines the residual connections with Inception v3 architecture, In other models adding more layers may lead to vanishing/ exploding gradient problems and also increases the testing as well as training error rates. Moreover, the increasing number of layers becomes difficult to train and the accuracy starts saturating, thereby the performance gets degrades. So the work uses Residual model, which efficiently solves such problems by using skip connections. Furthermore, the memory of the classifier is retained by the long short-term memory (LSTM), which efficiently handles the memory rates of the classifier.
The softmax layer is removed and the features are passed to LSTM to predict the abnormal activity.

Finally, the key points observed in each model are: VGG16-Visual Geometry Group 16 is trained up to 22,000 categories and the filter size is uniform throughout all the layers. The drawback is, it is slow to train and the network architecture weights are quite large. ResNet called Residual Network, the strengths of ResNet is, it takes less memory, Faster Inference Time, Allows deeper networks to be trained, supports skip connections, which help to increase performance. The weakness is increased complexity of architecture, Implementation of Batch Normalization is needed, Adding skip connections for which taken into account the dimensionality between the different layers which can become a headache to manage. It's a 50 layers architecture and can be trained up to 23 million trainable parameters. One of the adv of 3DCNN captures better spatio-temporal info. in videos and the weakness is it takes more time for execution. Inception V3 important features are, it has less parameters than VGG16. Inception V3 model learns more complex features, its computationally efficient and auxiliary classifier are used. Training the network is much faster and has better final accuracy in Recurrent-Residual-Inception V3.

3. RESULTS

The performance comparison of the six deep learning based LSTM models are done using following setup

Table 1: Performance Configuration

PC configuration	Intel i7, 8 GB RAM, Nvidia MX350
Software tools	Python 3.7, Keras, tensor flow
Dataset	UCF-Crime dataset
Number of Training samples	800 normal videos, 810 anomalous videos
Number of Test samples	150 normal, 140 anomalous Videos

The train and test accuracy over various epochs for Resnet, VGG16, VGG19, 3DCNN, Inception V3, Recurrent-Residual-Inception V3 with LSTM is given below

Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Comparison of Loss And Accuracy Before Pre-Processing

Model	Training		Test	
	Loss	Accuracy	Loss	Accuracy
Resnet with LSTM	0.49	0.73	0.60	0.70
VGG16 with LSTM	0.64	0.70	0.60	0.70
VGG19 with LSTM	0.64	0.60	0.68	0.72
3DCNN with LSTM	0.60	0.68	0.54	0.69
Inception V3 with LSTM	0.66	0.75	0.58	0.70
Recurrent-Residual-Inception V3	0.42	0.80	0.47	0.75

The comparison of training and test accuracy and loss across the six models are the Resnet, Inception V3 and Recurrent-Residual-Inception V3 with LSTM are able to achieve an accuracy of 70% at epoch of 4 seconds. VGG19 with LSTM is able to achieve 72% accuracy only at epoch of 8 seconds. RRIV3 with LSTM is able to achieve the highest test accuracy of 76% at epoch of 7 seconds. Out of all methods, RRIV3 with LSTM is able to achieve good training and test accuracy but it is observed that the loss is more in all the models except RRIV3. Use of residual layer and better protection against local gradient has helped RRIV3 with LSTM to achieve the highest accuracy in fewer epochs compared to others.

The loss achieves it minimal value of 42% at epoch of 4 seconds in RRIV3 with LSTM. VGG16 with LSTM is able achieve loss of 64% at 9 seconds. VGG19 with LSTM is able to achieve loss of 64% at 9 seconds. Loss is only 60% in Inception V3 with LSTM at epoch of 6 seconds. RRIV3 with LSTM is able to achieve a loss of 42% at epoch of 1 second. The loss is lowest in RRIV3 with LSTM due to use of convolutions of different sizes to capture details at varied scales.

The metrics used are for measuring and comparing the effectiveness are standard metrics of accuracy, precision, recall and execution time.

The comparison of training and test accuracy and loss across the six models after applying preprocessing are given below

Table 3: Comparison of Loss And Accuracy After Pre-Processing

Model	Training		Test	
	Loss	Accuracy	Loss	Accuracy
Resnet with LSTM	0.34	0.83	0.45	0.80
VGG16 with LSTM	0.54	0.80	0.50	0.79
VGG19 with LSTM	0.45	0.82	0.54	0.80
3DCNN with LSTM	0.46	0.76	0.44	0.79
Inception V3 with LSTM	0.42	0.83	0.48	0.80
Recurrent - Residual-Inception V3	0.22	0.90	0.27	0.85

The result shows that two models of Resnet, Inception V3 with LSTM are able to achieve same test accuracy of 83%. The accuracy of 3DCNN is 76%, which is lower compared to RRIV3 model. Therefore, it is observed that loss is very low in RRIV3 with LSTM at 22% training and 0.27% for test.

The average execution time is compared against all the six different models and the result is given below.

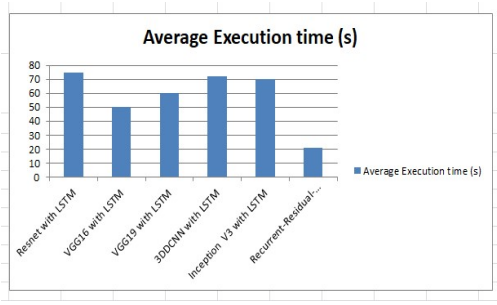

Figure 3: Execution Time of Models

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The execution time is lower in RRIV3 with LSTM compared to other models due to residual connections. 3DCNN feature extraction takes almost same execution time as that of Resnet with LSTM even though layers are less due to 3D feature complexity.

The precision for detection of abnormal events in each of the models is given below

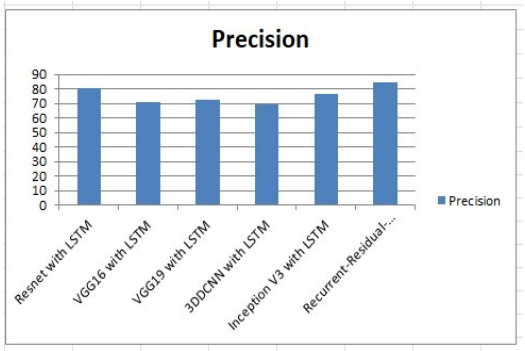

Figure 4: Precision Across Models

The recall for detection of abnormal events in each of the models is given below

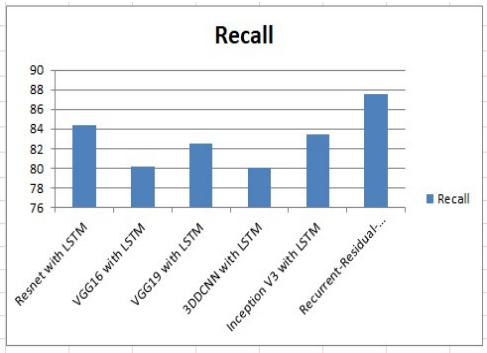

Figure 5: Recall Across Models

As per the results shown in the Figure 4 and 5 our suggested method, outperformed significantly well compared to other models.

The main goal of early action prediction is to identify partially observed videos, which does not provide sufficient information either for recognition or prediction of actions beforehand. Thus early action prediction is a challenging task. The below output shows the step by step process in figure 6 and 7 and final output in figure 8..

Experiments were carried out on UCF-Crime dataset and below are the frames extracted for a sample Burglary video in UCF-Crime Dataset.

Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 6: Extracted Frames from Burglary video Dataset.

Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

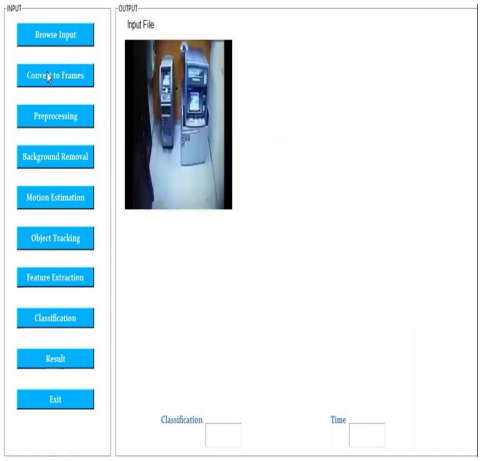

Figure 7: (a)

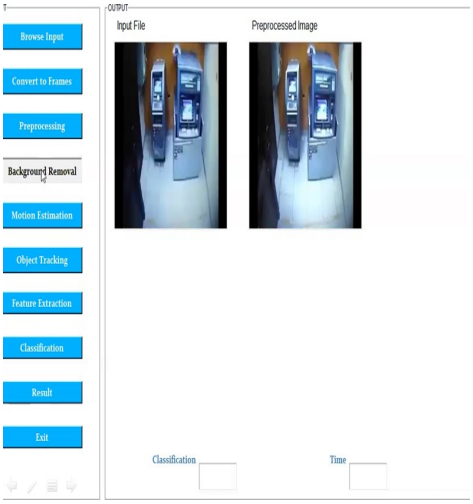

Figure 7: (b)

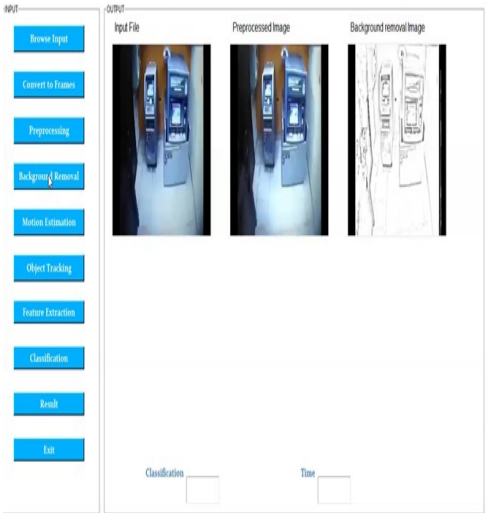

Figure 7: (c)

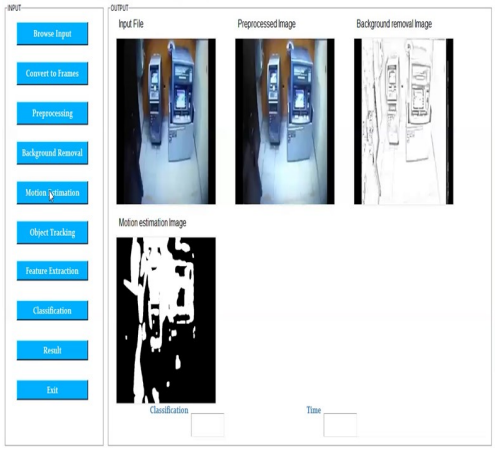

Figure 7: (d)

Figure 7: (e)

Figure 7: (a) *Video File Is Selected As An Input And Converted In To Frames. (b) Preprocessing On Frames Is Done Using Histogram Partition Based Gamma Correction Method.(c) Background Removal Using Modified Livewire Segmentation. (d) Resulted Output After Applying Applied Motion Estimation. (e) Result After Applying Yolov4*

The below figure shows the output after applying different steps on the video frames and finally produces the predicted frame as an output.

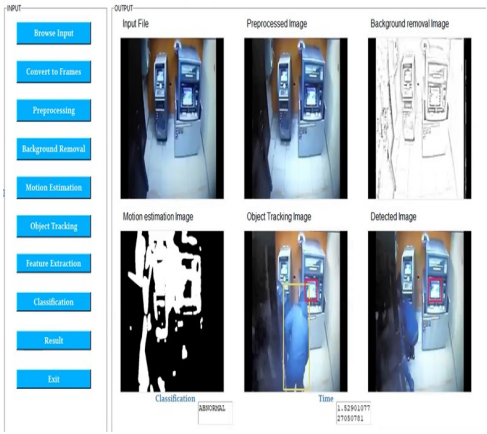

Figure 8: Early Action Predicted Frame Is The Final Output

4. CONCLUSION

The effectiveness of six existing different deep learning models are compared and proposed a better method for solving the luminance and shadow problem in video surveillance system that helps in abnormal activity prediction accurately. VGG19, VGG16, 3DCNN, Inception V3 and Resnet when combined with LSTM were able to provide an accuracy of 80% where as Recurrent-Residual-Inception V3 (RRIV3) with LSTM was able to provide training accuracy of 90%, a lowest loss of 22% with a lower execution time of 20.6 microseconds compared to other deep learning with LSTM models. Recurrent-Residual-InceptionV3 network (RRIV3) is used for reducing the execution time and makes the network wider as well as deeper to reduce the load during the training phase as well as for increasing efficiency. Inception block gets more data by the varying scales of input frames. Residual network gains accuracy because of increased depth hence added recursive nature to residual block.

RRIV3 outperformed in terms of accuracy, it takes less execution time and the performance is based on recognizing the actions and predicting the actions beforehand. Experiments were carried out on UCF-Crime dataset. RRIV3 performed well on test dataset.

The work can be extended further to larger classes of abnormal event and instead of sequential frame, observational frames can be passed.

REFERENCES

[1] Y. Li, Y. Cai, J. Liu, S. Lang, X. Zhang, "Spatio-temporal unity networking for video anomaly detection", IEEE Access 7 (2019) 172425–172432.

[2] T. Wang, M. Qiao, Z. Lin, C. Li, H. Snoussi, Z. Liu, C. Choi, "Generative neural networks for anomaly detection in crowded scenes", IEEE Transcations on Information Forensics and Security 14 (5) (2018) 1390–1399

[3] Lee, Sangmin et al. "Stan: Spatio- Temporal Adversarial Networks for Abnormal Event Detection." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018): 1323-1327.

[4] Athanesious, Joshan; Srinivasan, Vasuhi; Vijayakumar, Vaidehi; Christobel, Shiny; Sethuraman, Sibi Chakkaravarthy: "Detecting abnormal events in traffic video surveillance using superorientation optical flow feature", IET Image Processing, 2020, 14, (9), p. 1881-1891.

[5] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, EnverSangineto, and Nicu Sebe. "Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection", eprint:1610.00307, arXiv,2016.

[6] Doshi, Keval, and Yilmaz, Yasin. "Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate". Pattern Recognition Volume 114, 2021,107865,

[7] Ullah, Waseem, et al. "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos." Sensors 21.8 (2021): 2811

[8] P. Singh and V. Pankajakshan, "A Deep Learning Based Technique for Anomaly Detection in Surveillance Videos," 2018 Twenty Fourth National Conference on Communications (NCC), 2018, pp. 1-6.

[9] Ko, Kwangeun & Sim, Kwee-Bo. (2018). "Deep convolutional framework for abnormal behavior detection in a smart surveillance system". Engineering Applications of Artificial Intelligence. 67. 226-234. 10.1016/j.engappai.2017.10.001.

[10] Majhi, Snehashis & Dash, Ratnakar & Sa, Pankaj. (2020). "Two-Stream CNN Architecture for Anomalous Event Detection in Real World Scenarios". 10.1007/978-981-15-4018-9_31.