J. Integr. Sci. Technol. 2024, 12(1), 704 Article

Journal of Integrated
SCIENCE & TECHNOLOGY

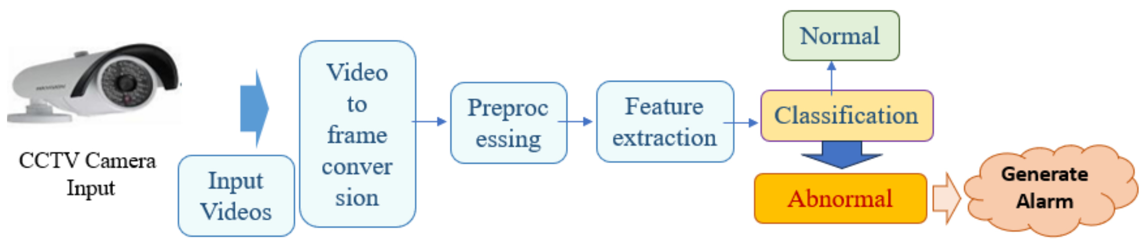
Human Anomalous Activity detection with CNN-LSTM approach

Megha G. Pallewar,[1*] Vijaya R. Pawar,[2] Arun N. Gaikwad[1]

[1]Zeal College of Engineering and Research, Pune, India, [2]Bharati Vidyapeeth's College of Engineering for Women, Pune, India.

Received on: 15-Jun-2023, Accepted and Published on: 05-Aug-2023

ABSTRACT



Human action recognition in digital media is a challenging task. Its processing has become popular recently due to increasing demand in fields related to human security. Machine learning based classification and recognition is the prominent approach for anomalous activity detection. The present work focuses on video anomalous activity detection. UCF crime database is used in the present work. Preprocessing operations are performed on videos from the dataset. The present system employs cascaded approach using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). The frames are applied to CNN as an input. CNN process the input and extract the prominent features. Convolution and pooling are the main operations in CNN. LSTM is used for the classification of the activities performed in the input video. Cascading of CNN and LSTM recognizes the six anomalous activities identified as Abuse, Arson, Assault, Burglary, Fight, and Robbery. 80% dataset is used for training, 10% for testing and 10% for validation and cross validation. The developed system attains approximately 99% accuracy and is a robust model.

Keywords: Anomalous activity detection, Convolutional neural networks, Long short time memory, Accuracy, Human activity detection

INTRODUCTION

Nowadays, globally video automated surveillance is a topic of concern and plays vital role for human security and also for national security. There is increasing demand for security in public places like bus stands, railway stations, airports, malls, cinema halls, schools and colleges, supermarkets and many more places of national importance.[1,2] The main objective of automated video surveillance is to detect the activities which can be categorised as normal and anomalous (abnormal) activities. The surveillance cameras are used to monitor all day to day activities at the public places and identify anomalous activity.[3,4] The primary objective of using surveillance cameras is to achieve early detection of anomalous human behavior.[5] This is vital in situations where human intervention is necessary for crime prevention and countering terrorism. However, manually monitoring surveillance footage can be labour-intensive and time-consuming, as abnormal events occur in only a tiny fraction of the time, while the majority of surveillance time remains uneventful. This leads to inefficiency and unnecessary waste of resources, including storage space for redundant video data. The anomaly can be categorized by different ways but the UCF Crime dataset has altogether 13 abnormal activities listed as, Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fight, Road Accident, Robbery, Shooting, Stealing, Shop lifting and Vandalism. An efficient surveillance system which is capable of automatically detecting strange behavior that may lead to dangerous situations is essential. To achieve this goal deep and comprehensive study of human activity recognition is essential. Through this study understanding of the distinctive features of each action is possible. Anomaly detection in video has broad applications, ranging from traffic accident detection to identifying illegal activities.[6]

Computer vision and pattern recognition are the emerging domains and can be used for interaction between objects and the subjects available in the videos. Convolutional Neural Networks (CNNs) is highly effective for anomalous activity detection in video surveillance and related applications.[7–9] CNN's excel at learning hierarchical features from raw data, making them

*Corresponding Author: Megha Pallewar
Email: msyannawar@gmail.com

Cite as: *J. Integr. Sci. Technol., 2024, 12(1), 704.*
URN:NBN:sciencein.jist.2024.v12.704

ISSN: 2321-4635
http://pubs.thesciencein.org/jist

particularly well-suited for visual tasks like image and video analysis.

Main objective of the present work is to develop the activity recognition and classification system from the captured video. CNN-Long Short Term Memory (LSTM) (CNN-LSTM)[10,11] model is employed for detection and recognition of anomalous activities.[12] CNN approach is used for feature extraction and LSTM is used for classification purpose.

Literature Survey

The work reported and national and international level is surveyed and cognizance of the work is presented here.

Xu et al. 2017[13] proposed a model which take dense points from each frame and these point are used further to develop dense trajectory approach and used for tracking object depending on the optical flow based information. The model used for multiple datasets and the maximum accuracy reported is 94.2% for KTH database and the minimum one is 58.3% for Hollywood database.

Liu et al., 2013[14] proposed system for identifying and recognizing different human activities. Optical flow graph along with a edge detector named Harris 3D is uesd to get spatial information from the video frames. Further Universal Background Model (UBM) is trained for calculation of local features such as Scale Invariant Feature Transform (SITF) and Spatio-Temporal Interest Points (STIP).

Bhalodia et al. 2010[15] proposed very significant concept based on deep learning which is nothing but development of attention mechanism that is utilized to extract relevant information from the focused area of a scene. The attention regions are then developed based on this information which can be the only focused input for intelligent surveillance systems.

Sharma et al. 2015[16] proposed a neural network LSTM model which works with attention mechanism for recognizing actions in videos. The implemented system tested with small datasets and only soft attention models are focused due to this computational cost is more.

Meng et al. 2019[17] implemented a mechanism which is easy to connect and interpret spatio-temporal attention mechanism. This majorly focuses on features of spatial domain utilized to apply on a convolutional LSTM mechanism for promising results. The developed model uses Intersection Over Union (IoU) for prediction and then compared with ground truth. Still the maximum video classification accuracy is 78.33%.

Singh et al 2020[18] designed a system for real time anomaly recognition using CNN and RNN model. A pre-trained InceptionV3 model is used for object recognition. Background based estimation and body based detection were performed to identify the violent and peaceful activity.

Huang et al 2020[10] developed a convolutional neural network called temporal spatial convolutional neural network (TSCNN) which takes continuous frame sequence as input. The network designed using 3D convolution to identify real time human activities. The RF-men dataset is created by using 6000 RFID signal data and converting that into pixel maps. The average classification accuracy is 94.6% and the lowest derived accuracy is 81.8%.

Tang et al 2021[19] proposed a lightweight CNN using Lego filters for human activity recognition (HAR). By using Lego CNN the problems of traditional CNN like processing units (filters) need temporal dimensions and sharing of units among multiple sensors are overcome. So by replacing the ordinary filters by Lego filters one can build more efficient HAR model. The established model then used on 5 public datasets and then their results are compared.

Roberta et al 2022[20] created Abnormal Activities Dataset which has eleven classes with 1069 videos. The proposed ConvLSTM architecture consists of ConvLSTM layer, a Conv2D layer and a time distributed layer. The results were then compared with already proposed architectures like 3D Resnet50, 3D Resnet101, and 3D Resnet152. The proposed system gives best results for created dataset. The achieved accuracy is 96.19%.

Literature survey reveals that an accuracy reported at national and international level for automatic video surveillance is high with defined and restricted parameters. So in the proposed system unrestricted parameter approach is used with machine learning based classifiers.

Present Methodology

Figure 1 depicts the methodology used for Video based Human Anamolous Activity Detection with Convolutional Neural Network and LSTM. It consists of dataset acquisition, extraction of videos and frames from dataset.

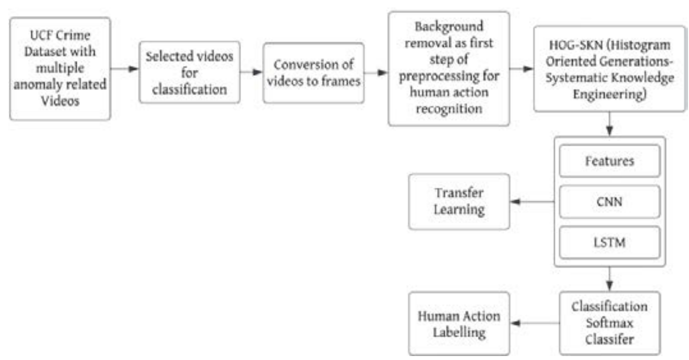


Figure 1. The present methodology used for anomalous activity detection using CNN and LSTM.

Database:

UCF crime dataset is used for research work. This dataset the consists of huge uncleaned surveillance videos which are with 13 most common real world crime videos of following categories: Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fight, Road Accident, Robbery, Shooting, Stealing, Shop lifting, Vandalism. This dataset contains around 1900 videos with each average frame size of 7247 with approximate 128 hours play time altogether. For our first phase of research our work considered with Abuse and Fighting classification and predictions. Further we used total six actions with same set of algorithms.

The flow of research work is as follows:

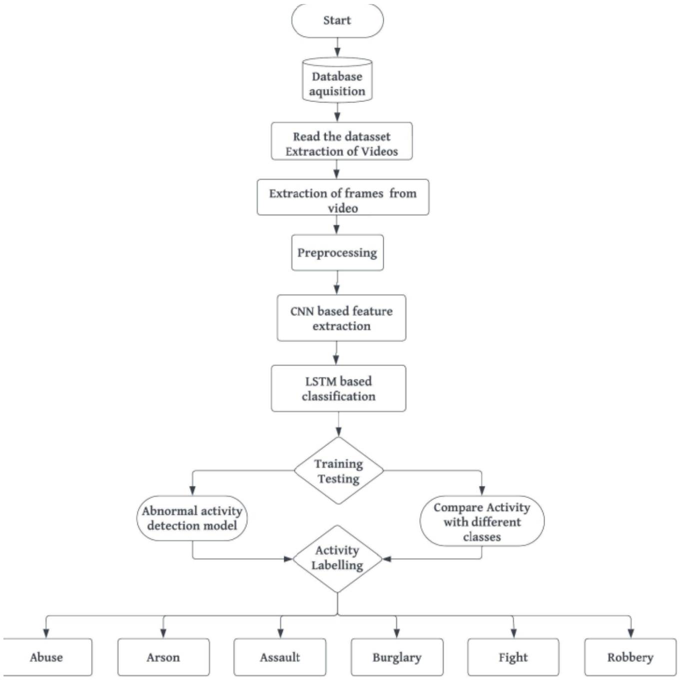


Figure 2. Flowchart of current system

LAYERED CNN-LSTM MODEL:

The detailed system architecture is depicted in Figure 1. Initially first step is to extract frames from videos with threshold of 40 frames to CNN-LSTM architecture and the approach is illustrated in figure 3. The input video is totally normalized into frames and 40 frames are used for further processing. CNN model is trained with 80% of videos from database. Trained CNN model is utilized to get the elements and features of the video. 25088 features are extracted from each frame. Total 40×25088 features are extracted and applied to LSTM as an input vector. LSTM engages these individual vectors from one stage and process them to the next stage. Finally the output comprises of dense layer model with Not Connected (NC) nodes.

CNN- CONVOLUTIONAL NEURAL NETWORK:

CNN is a dense network of layers. Each layer is dependent on the output of its prior layer. A sequence of convolutional layer, sub sampling layer and dense layer makes the convolutional neural network layers. Convolutional layer will be having the k kernels or filters deployed to produce k feature maps as vectors. In sub sampling layer every feature vector further sampled that is named as max pooling operation, which effectively reduces the spatial size used for representation. The number of attributes which are trained will be also reduced due to max pooling. The dense layer with fully interconnected layers then used to classify the image frame into a match out of many classes available in dataset. The detailed CNN operation in pre-trained mode is as shown in Figure 4.

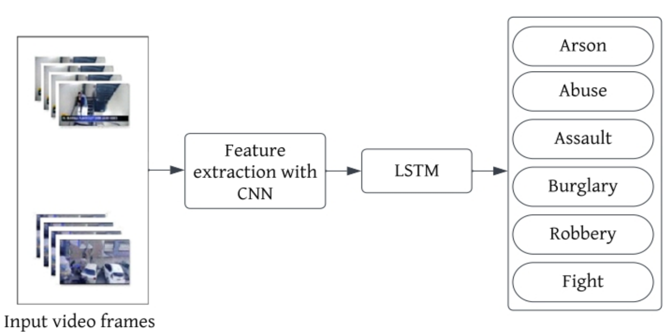


Figure 3. System Architecture

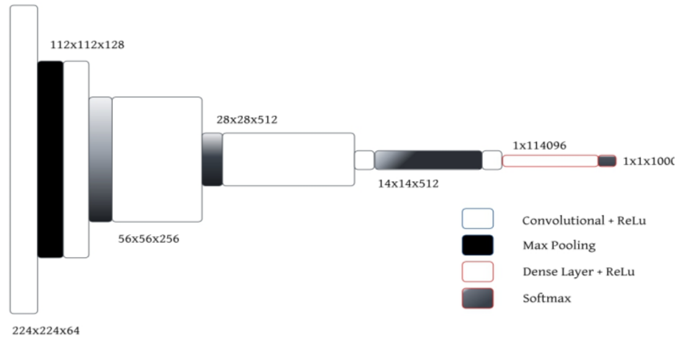


Figure 4. Pre-trained CNN implemented in Keras

The input image size is kept fixed as 224x224x3. Then the input is forwarded as a sequence convolutional layer like a stack which uses ReLu as an activation function. The activation function used has each of 4096 nodes. CNN is consist of max pooling operation which reduces the size of output feature vectors. The dense layer uses Softmax as an activation function. The final output of convolutional neural network is carried out by approximately 1500 nodes. Figure.5 shows model summary of implemented CNN.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
rescaling_1 (Rescaling)      (None, 180, 180, 3)       0
_________________________________________________________________
conv2d (Conv2D)              (None, 180, 180, 16)      448
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 90, 90, 16)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 90, 90, 32)        4640
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 45, 45, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 45, 45, 64)        18496
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 22, 22, 64)        0
_________________________________________________________________
flatten (Flatten)            (None, 30976)             0
_________________________________________________________________
dense (Dense)                (None, 128)               3965056
_________________________________________________________________
dense_1 (Dense)              (None, 12)                1548
=================================================================
Total params: 3,990,188
Trainable params: 3,990,188
Non-trainable params: 0
```

Figure 5. Model summary of CNN

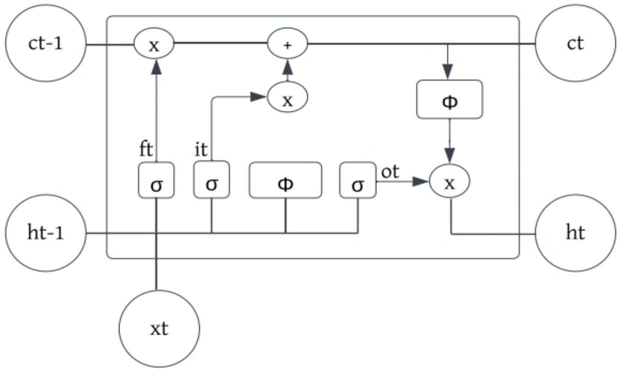


Figure 6. LSTM Model

LSTM (Long Short-Term Memory):

LSTM is a type of Recurrent Neural Network. The LSTM typical model is shown in Figure 6. It is naturally formed in the way that it remembers the information for long time period. The various time steps included are: x_t is input, h_{t-1} is previous output, c_{t-1} is previous memory unit. And the associated outputs are: h_t is current output and c_t is current memory unit.

Generally, LSTM is designed with three types of logical GATE which is the building blocks of the system. The h_t and c_t are the important element of LSTM. The two units (h_t and c_t) are analysed by following equations:

1. Input gate i: Handles the current entry x_t .

   $f_t=\sigma\ (W_{xf}x_t + W_{hf}\, h_{t-1} + b_f)$………………..(1)

2. Forget gate f: Manages the LSTM to forget the preceding memory c_{t-1}.

   $ot = \sigma\ (Wwoxt + Whoht\text{-}1 + bo)$…………..(2)

3. Output gate o: Decides what amount of memory to be transferred to the hidden state h_t .

   $o_t = \sigma\ (W_{xo}x_t + W_{ho}h_{\ t-1} + b_o)$……………...(3)

4. Finally:

   $c_t = f_t \otimes c_{t-1} \oplus i_t \otimes \varphi(W_{xc}x_t + W_{hc}h_t\text{-}1 + b_c)$
   $h_t = o_t \otimes \varphi\ (c_t)$…………………………….(4)

Where: σ is the sigmoidal function, φ is the hyperbolic tangent, ⊗ represents the product with the value of the gate and the weights of the matrix denoted by W_{ij}.

Figure 7 and Table 1a shows the process of LSTM algorithm and LSTM model summary respectively.

Experimentation and Results

The results are obtained from the testing with 20, 40 and 60 frames after training performed on 40 frames for each and individual action class. While performing the experimentation work started with 10, 20 and then 30 epochs along with variable batch size, filter and dense layer count. Total 30 epochs are applied to get these results. Here our work trained for firstly two, then six human action classifications which lead us to build models to identify abnormal activity like Abuse, Arson, Assault, Burglary, Fight, and Robbery.

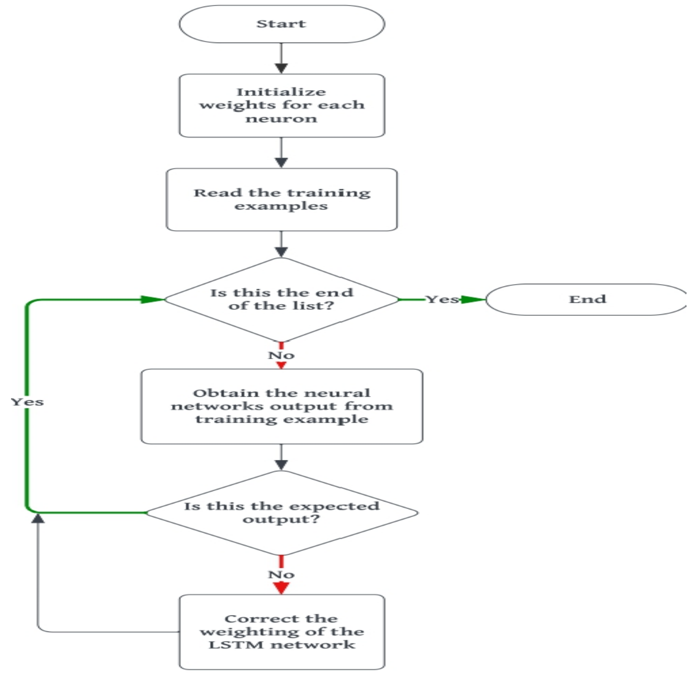


Figure 7. Algorithm flowchart of implemented LSTM model

Table 1a. LSTM Summary

Layer (Type)	Output Shape	Number of Parameters
Input Layer	(None, 30, 1000)	0
LSTM	(None, 30, 256)	1,28,7168
LSTM	(None, 30, 128)	197,120
LSTM	(None, 30, 64)	49,408
Flatten	(None, 1920)	0
Dense	(None, 256)	491,776
Dropout	(None, 256)	0
Dense	(None, 128)	32,896
Dropout	(None, 128)	0
Dense	(None, 2)	258

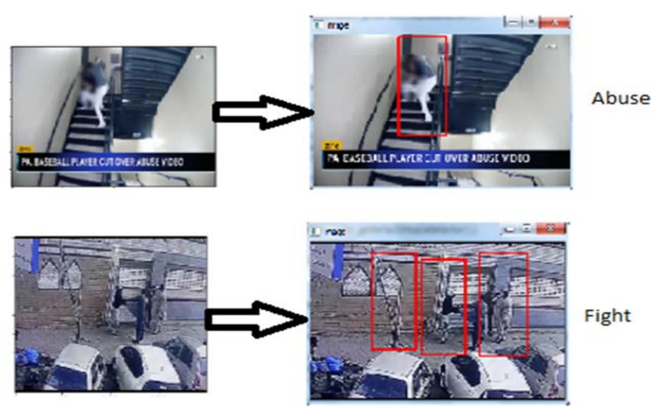


Figure 8. Actual prediction of actions of Abuse and Fight with human recognition with boundaries

Figure 8 and 9 shows actual prediction for 2 activities on testing with non-trained input frames with 100% accuracy.

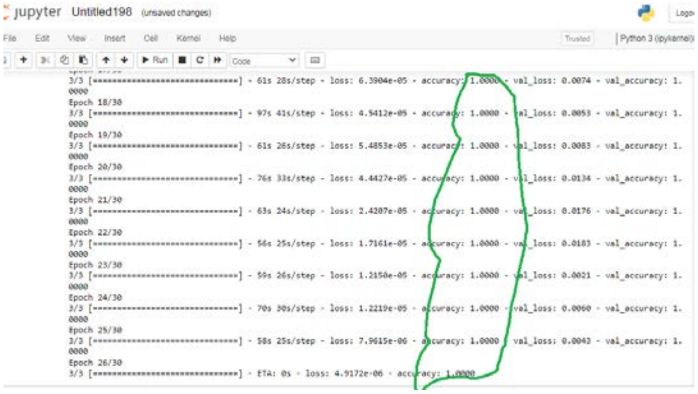


Figure 9. Training with 30 epoch showing 100% accuracy

The CNN-LSTM model is then trained for random 6 activity classes. The model performed really well for detecting the anomalous activity in the scene of interest.

```
15384
Found 15384 files belonging to 6 classes.
Using 12308 files for training.
Found 15384 files belonging to 6 classes.
Using 3076 files for validation.
['Arson', 'Assault', 'Burglary', 'Robbery', 'abuse', 'fight'
```

Figure 10. Class result window from Jupyter notebook

The model worked perfectly for six classes and the results are shown in figure 10, 11 and 12. Figure 10 shows 6 class results and figure 11 labelling done by model for all six classes. Figure 13 shows 100% accuracy achieved after only 10 epochs with CNN-LSTM model.

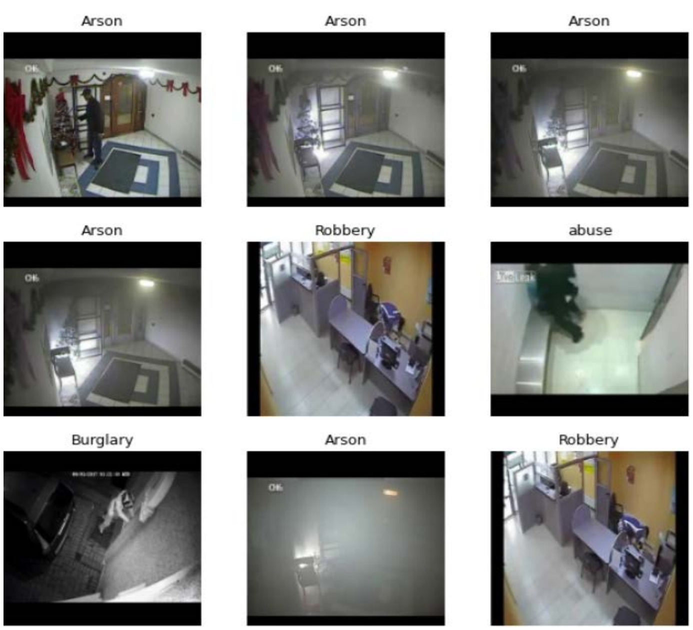


Figure 11. Six activity labeling

Figure 14 shows the grid pattern display for 6 activity classes where all activities are identified for 20 sample frames. The model tested similarly for 40, 60, 80 and 100 frames of every class.

Figure 12. Result for six activity testing

Performance Analysis And Discussion

Table 1b shows the performance analysis of the present and existing systems with respect to dataset complexity. The dataset duration, location of videos, view in scene, application, actions undertaken and average accuracy using the mentioned dataset is discussed in this table. The data presented elaborate the complexity of used dataset and accuracy obtained highlights the best performance of the present system.

Table 1b. Performance of present and existing systems with respect to dataset used.

Type of database	Video recording details	Scene Type	Area of application	Actions under taken	Accuracy obtained (%)
UCF Crime dataset (The present system dataset)	Total duration≈ 128 hrs Total no. of frames ≈ 13 million Total no. of videos ≈ 950(N)+950(AN)	Single-scene and Multi-scene Indoor Outdoor videos	Public monitoring Security -threat	abuse, arrest, assault, arson, accident, burglary, fight, shoplifting, vandalism, robbery, explosion, shooting, stealing	≈100
USCD Ped1	Total duration≈ ≈ 5-7 min Total no. of frames ≈ 14000 Total no. of videos ≈ 34(N)+36(AN)	Single-scene Outdoor videos	Public monitoring Non Security -threat	the circulation of non-pedestrian entities in the walkway (bikers, skaters, small carts)	≈92
USCD Ped2	Total duration≈ 2-3 min Total no. of frames ≈ 4560 Total no. of videos ≈ 16(N)+14(AN)	Single-scene Outdoor videos	Public monitoring Non Security -threat	anomalous pedestrian motion patterns	≈96

Shanghai iTech	Total duration≈ ≈ 3-3.5 hrs Total no. of frames ≈ 317398 Total no. of videos ≈ 330(N)+107(AN)	Multi-scene Outdoor videos	Public monitoring Security -Non security threat	throwing objects, loitering and running, walking in wrong directions and loitering	≈85.94
Street Scene	Total duration≈ 3-4 h Total no. of frames ≈ 203257 Total no. of videos ≈ 46(N)+35(AN)	Single-scene Outdoor videos	Not specific application	Cars driving, turning, stopping and parking; pedestrians walking, jogging and pushing strollers; and bikers riding in bike lanes	≈94.2
Avenue	Total duration≈ 20 min 26 sec Total no. of frames ≈ 30652 Total no. of videos ≈ 16(N)+21(AN)	Single-scene Outdoor videos	Public monitoring Security -Non security threat	Strange action, Wrong direction, Abnormal object	≈92.3
HTA	Total duration ≈ 4 hrs Total no. of frames ≈ 0.4 million Total no. of videos ≈ 286(N)+107(AN)	Multi-scene Outdoor videos	Traffic monitoring	Abnormal traffic videos on highways	≈82.8
UMN	Total duration≈ 4 min 17 sec Total no. of frames ≈ 7710 Total no. of videos ≈ 11	Single-scene and Multi-scene Indoor Outdoor videos	Public monitoring Security -threat	Crowd Escape Panic	≈98.8

Table 2 depicts that existing models i.e. 3D-ConvNet architecture, OFG with Harries 3D edge detector and RNN LSTM model attained 94.39%, 93.67% and 95.20% accuracy respectively. The present system attains 100% accuracy which is maximum accuracy for anomaly detection.

Table 2. Result of Video Classification

Model used	Accuracy (%)
3D-ConvNet architecture	94.39
OFG with Harries 3D edge detector	93.67
RNN LSTM model	95.20
CNN-LSTM (Our model)	100

The environmental setup used to implement the present system is shown in Table 3.

Table 3. Result of Video Classification

Machine Configuration	2.7 GHz Intel Core i7-4600U CPU 8GB RAM 64 bit Operating system
Environment	Python 3.7.0
Dataset	UCF crime dataset

Conclusion and Future Scope

Through the experimental results it can be said that the present work performs perfectly for UCF crime dataset. The CNN-LSTM combination model shows efficient results for range of anomalous activities that has maximum chances in general locality. For human safety the proposed system can be utilized and the achieved output then can be fed to any indicator device or system or to any authorised firm which can take action against this abnormal behaviour in communal place. In further work we will be working with all 13 classes of UCF crime dataset and then it is quite possible that we can achieve the similar results for very first time. The training of the model takes too much time which is only drawback of our model. But that can be also solved by focusing on only part of scene where activity observed and ignoring the other part of the frame. This will automatically effect on the response time of the system and so the performance parameters would definitely improve.

Conflict of Interest: The authors declare that they have no conflict of interest.

Acknowledgments

The authors appreciate Mr. Sham Yannawar and Mr. Ganesh Pallewar for their continous motivation and invaluable help, guidance throughout this work. Authors express gratitude to Mr. M.G. Unde, Dean R&D and Dr. A.M. Kate, Principal ZCOER for providing the necessary facilities as well as kind support.

References

1. M. Helmy, W. El-Shafai, S. El-Rabaie, I.M. El-Dokany, F.E.A. El-Samie. Efficient security framework for reliable wireless 3D video transmission; Springer US, **2022**; Vol. 33.
2. H.T. Duong, V.T. Le, V.T. Hoang. Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey. *Sensors* **2023**, 23 (11), 5024.
3. A. Hendel, D. Weinshall, S. Peleg. Identifying surprising events in video using bayesian topic models. *Stud. Comput. Intell.* **2012**, 384, 97–105.
4. S.S. Priya, R.I. Minu. Abnormal Activity Detection Techniques in Intelligent Video Surveillance: A Survey. In *7th International Conference on Trends in Electronics and Informatics, ICOEI 2023 - Proceedings*; **2023**; pp 1608–1613.
5. S.N. Kakarwal, A.S. Gavali. Context aware human activity prediction in videos using Hand-Centric features and Dynamic programming based prediction algorithm. *J. Integr. Sci. Technol.* **2022**, 10 (1), 11–17.
6. J. Ren, F. Xia, Y. Liu, I. Lee. Deep Video Anomaly Detection: Opportunities and Challenges. *IEEE Int. Conf. Data Min. Work. ICDMW* **2021**, 2021-Decem, 959–966.
7. B. Kabra, C. Nagar. Convolutional Neural Network based sentiment analysis with TF-IDF based vectorization. *J. Integr. Sci. Technol.* **2023**, 11 (3), 503.
8. S. Maurya, S. Raj, N.R. Bhoi, et al. Morphological analysis of metabolically dysregulated spermatozoa using Artificial Intelligence based approach. *J. Integr. Sci. Technol.* **2023**, 11 (4), 569.
9. R.R. Karwa, S.R. Gupta. Automated hybrid Deep Neural Network model for fake news identification and classification in social networks. *J. Integr. Sci. Technol.* **2022**, 10 (2), 110–119.
10. W. Huang, Y. Liu, S. Zhu, S. Wang, Y. Zhang. TSCNN: A 3D Convolutional Activity Recognition Network Based on RFID RSSI. In *Proceedings of the International Joint Conference on Neural Networks*; **2020**.
11. D. Zhang, W. Wu, H. Cheng, et al. Image-to-Video Person Re-Identification with Temporally Memorized Similarity Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, 28 (10), 2622–2632.

12. Y. Parikh, H. Koringa. Left Ventricle Segmentation using Bidirectional Convolution Dense Unet. *J. Integr. Sci. Technol.* **2023**, 11 (1), 417.
13. P. Xu, Z. Yu, W. Jin, H. Jiang. Action Recognition by Improved Dense Trajectories. In *Xitong Fangzhen Xuebao / Journal of System Simulation*; Colorado Springs, United States, **2017**; Vol. 29, pp 2053–2058.
14. Y. Cong, J. Yuan, J. Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognit.* **2013**, 46 (7), 1851–1864.
15. V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos. Anomaly detection in crowded scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; **2010**; pp 1975–1981.
16. S. Sharma, R. Kiros, R. Salakhutdinov. Action Recognition using Visual Attention. *CVPR* **2015**, 271–350.
17. L. Meng, B. Zhao, B. Chang, et al. Interpretable spatio-temporal attention for video action recognition. In *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*; **2019**; pp 1513–1522.
18. V. Singh, S. Singh, P. Gupta. Real-Time Anomaly Recognition Through CCTV Using Neural Networks. *Procedia Comput. Sci.* **2020**, 173, 254–263.
19. Y. Tang, Q. Teng, L. Zhang, F. Min, J. He. Layer-Wise Training Convolutional Neural Networks with Smaller Filters for Human Activity Recognition Using Wearable Sensors. *IEEE Sens. J.* **2021**, 21 (1), 581–592.
20. R. Vrskova, R. Hudec, P. Kamencay, P. Sykora. A New Approach for Abnormal Human Activities Recognition Based on ConvLSTM Architecture. *Sensors* **2022**, 22 (8).

Authors details:

Megha G. Pallewar
Research Scholar,
ZCOER, Pune
Email ID:
meghapallewar21@gmail.com

Dr. Vijaya R. Pawar
Bharati Vidyapeeth's COEW
Email ID: pawarvr74@gmail.com

Dr. Arun N. Gaikwad
Research Guide, Former Principal,
ZCOER, Pune
Email ID: arungkwd47@gmail.com