Anomaly Detection in Video Surveillance using SlowFast Resnet-50

Mahasweta Joshi[1]
Computer Engineering Department
CSPIT, CHARUSAT University, Changa, India

Jitendra Chaudhari[2]
Electronics and Communication Department
CSPIT, CHARUSAT University, Changa, India

***Abstract*—Surveillance systems are widely used in malls, colleges, schools, shopping centers, airports, etc. This could be due to the increasing crime rate in daily life. It is a very tedious task to monitor and detect abnormal activities 24x7 from the surveillance system. So the detection of abnormal events from videos is a hugely demanding area of research. In this paper, the proposed framework is used for deep learning concepts. Here SlowFast Resnet50 has been used to extract and process the features. After that, the deep neural network has been applied to generate a class using the Softmax function. The proposed framework has been applied to the UCF-Crime dataset using Graphics Processing Unit (GPU). It includes 1900 videos with 13 classes. Our proposed algorithm is evaluated by accuracy. Our proposed algorithm works better than the existing algorithm. It achieves 47.8% more accuracy than state of art method and also achieves good accuracy compared to other approaches used for detecting abnormal activity on the UCF-Crime dataset.**

***Keywords*—Accuracy; GPU (Graphics Processing Unit); SlowFast Resnet50; Softmax; UCF-Crime dataset**

I. Introduction

Abnormal activity detection is a very monotonous process for monitoring and identifying the abnormal events. In daily life, the use of surveillance cameras is increasing due to the huge crime rate. It may happen for somebody activity is suspicious activity while for others it is not. So to choose an appropriate framework for identifying suspicious activity plays a very significant role [1]. Recently surveillance system uses deep learning models to identify abnormal activities. Deep learning is very popular and one of the classes of machine learning. It is popular because of its ability to perform well on unstructured data. Deep learning uses and implements deep learning neural networks. It has good computing capability. It also provides good flexibility when it has to process a large number of features. These features have been taken from unstructured data. The Deep learning models have several layers and it passes the data through these layers. Each layer extracts the features progressively from unstructured data. These features have been passed to the next layer of the network. Low-level features have been extracted from initial layers and succeeding layers combines the features. It creates a complete form of feature representation. It addresses a variety of challenges that occur in a conventional surveillance system. It gives better performance with deep neural networks [2]. The literature study [2, 3, 4, 5, 6] summarizes the following point that inspire to use of deep learning in the surveillance systems.

- Anomaly events are irregular and unknown.
- Due to this one class of anomaly may differ from the other class of anomaly.
- The scalability of the data also plays a vital role in increasing the complexities of a system.
- Beacause the chances of misclassification of anomalies is too costly than the normal instances, the classes are imbalanced.
- Diversity in types of abnormalities.

II. Related Work

Abnormal activity detection can be performed in two kinds of categories. The first one is an individual activity video and the second one is a crowd activity video. Sonkar et al. [7] represent the deep learning system which provides control over the crowd. This will help to avoid suspicious activities. CNN model has been used to analyze crowd behavior. With CNN model KNN- K Nearest Neighbour is also used to calculate the position difference between two consecutive frames of an object. The motion has been analyzed using three attributes: Speed, Direction and angle. Using the thresholding technique, the event is classified. This may vary based on the application. Chaudhary et al. [8] proposed framework, in which five different features are extracted. The features are Speed, Centroid, Direction, Movement and Dimensions. After extracting features, rule based classification has been performed. This algorithm is capable of handling multiple activities which are abnormal activities. It performs successfully and gives an accuracy up to 90%. Kaminski et al. [9] proposed an unsupervised method for detecting abnormality from videos. UMN dataset has been used. They have used motion descriptors for classification, Particle filter algorithm is used. This algorithm does not require any predefined samples. The results obtained are highly efficient compared to other existing algorithms. Landi et al. [10] explored the need for spatiotemporal tubes instead of the whole frame from the surveillance video. The experimental results show that the network trained with these features gives much better performance than the ordinary way. This method is robust against different localization errors. Dubey et al. [11] focuses on minimizing false rate in abnormal activity detection.

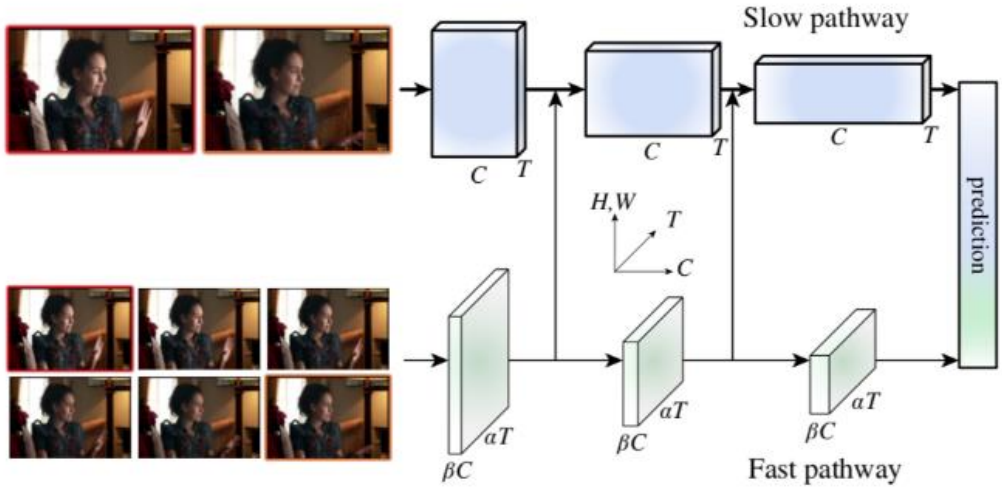


Fig. 1. Illustration of the SlowFast Network with Parameters [15].

3D Resnet has been used to extract spatio-temporal features. These features are used for deep neural networks and greatened scores. Multiple Instance Learning is used to classify the activities. The algorithm gives the best performance in the UCF-Crime dataset. Sargano et al. [12] present a novel method using CNN and a combination of SVM and KNN. In terms of accuracy, the proposed algorithm gives better performance than the existing methods. Sultani et al. [13] proposed a ranking method to classify the activities from the bag of videos. For better localization, sparsity and smoothness have been used. The results of experiments show that the proposed algorithm performs significantly higher than the other baseline algorithms.

III. SlowFast Resnet – 50

Generally, with videos, the frames consist of two part: static area in frames and Dynamic area in frames. The static area cannot be changed or slowly changed though the frame has been changed while in a dynamic area, it has changed the object, background and other things. For example, during a meeting, two persons doing handshaking is dynamic and fast but the background and other objects are static. Here in this paper, SlowFast CNN [14] has been introduced for capturing abnormality. Here slow pathway is designed to capture static information from video with low frame rates and slow refreshing speed. Another pathway is the fast pathway where all dynamic information is captured with high frame rates and fast refreshing speed. The formal pathway is very light weighted. Both pathway are merged by lateral connections. SlowFast network uses the Resnet model in both pathways and runs 3D convolution operations on it. The slow pathway uses a large strides. The stride means the number of frames skipped per second. In general, it is set to 16. Approximately allowing two sampled frames per second. The fast pathway uses too smaller a stride typically eight. This allows 15 frames per second.

Fig. 1 shows the illustration of the network with parameters. The parameters used in networks are as follows [15]:

- S = Spatial.
- T = Temporal.
- C = Channel size.
- α = Speed ratio (Frame skipping rate).
- β = Channel ratio.

Both the SlowFast network uses the 3D Resnet model. It captures several frames from videos and applies 3D convolution operation on them. The Slow Pathway uses a large stride, here it is 16. The Fast Pathway uses small size stride, here it uses two. As shown in Fig. 1, data from the Fast pathway faded to the Slow pathway via lateral connections. This connection allows the Slow pathway to be aware of the result of the Fast pathway. Data transformation should require as the shape of the data is different. This can be done by applying Time-strided convolution. In this, it performs 3D convolution of 5x12 kernel. Global average pooling has been one at the end of each pathway. It reduced its dimensionality of it. Now it concatenates the result of both pathways and inserts it into a fully connected classification layer [15].

IV. Proposed Method

Fig. 2 shows the flow of the proposed algorithm. The proposed algorithm has been classified as abnormal and normal activity. Here UCF – Crime data set has been used. It is a very large dataset and it contains 1900 videos. It includes videos of normal and abnormal activities. There are 13 abnormal activities in this dataset [13]. This algorithm has been implemented with the SlowFast Resnet50 model. Resnet is a residual network that implemented identity mapping. Because of it, Resnet is able to solve vanishing gradient problem. So in the proposed algorithm Resnet has been used. This model is pre trained model. The steps of the algorithm have been mentioned below:

1) Take a surveillance video S for processing.

2) Initialize the value of the number of frames and the number of the segment.

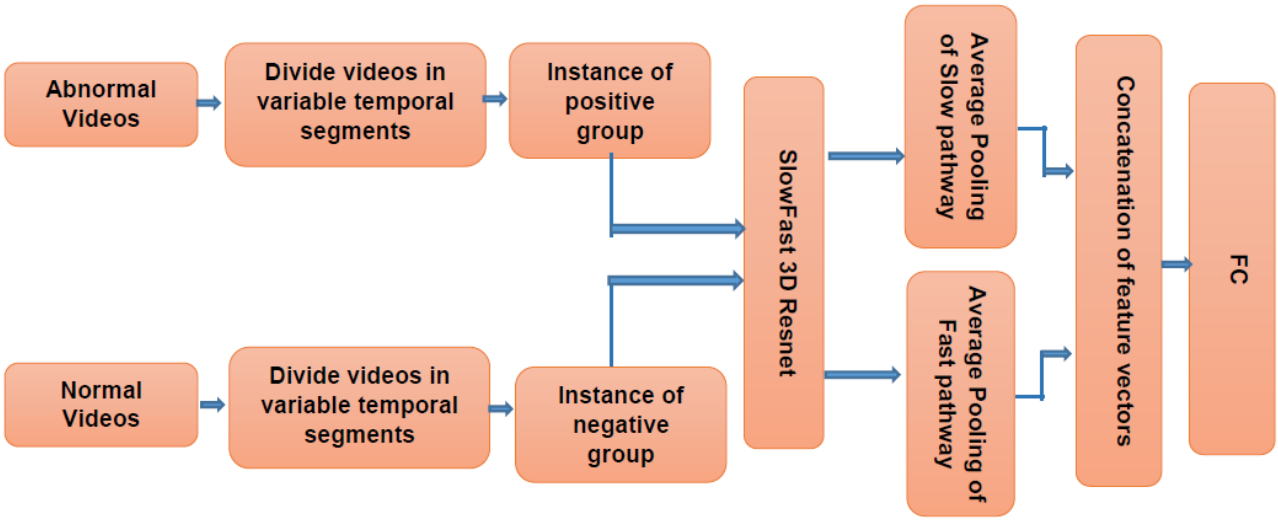


Fig. 2. Proposed Algorithm: Abnormal Activity Detection using SlowFast Resnet50 (AADSFR50).

3) Evenly divide Normal and abnormal videos into variable segments using the below formula.

$$clip = \frac{Total\_frame\_video}{I\_frame * I\_segment} \quad (1)$$

Where Total_frame_video is a total number of frames of the video, I_frame is the initial value of the number of frames and I_segment is the initial value of the segment.

4) The instance of normal and abnormal video groups has been given to SlowFast Resnet pre-trained models.

5) Data from the fast pathway is going to feed into the slow pathway network using the lateral connection.

6) Using average pooling, slow pathway feature vector has been generated.

7) Using average pooling, fast pathway feature vector has been generated.

8) Two feature vectors have been fused by performing concatenation.

9) A fused feature vector has been given to a fully connected classifier layer where the Softmax function is used to classify the activity.

V. Implementation Results and Discussion

UCF–Crime dataset has been used to evaluate the AADSFR50 – proposed algorithm. This dataset is widely used to research abnormal activity detection. It contains 1900 videos and 13 classes. The classes are Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism [13]. Google colab has been used to implement this algorithm. It provides NVIDIA Tesla K80.

Table I shows the comparison of different existing algorithms and our proposed algorithm (AADSFR50). Here as a measuring parameter Accuracy has been used. As observed from Table I, our proposed algorithm works better than the existing algorithm.

Fig. 3 describes the pictorial information of a comparative study of existing algorithms and our proposed algorithm AADSFR50. Here the evaluation parameter taken is accuracy. Fig. 3 shows that our proposed method works better than existing algorithms. It achieves 47.8% increase in accuracy compared to state of the art method. Other approaches also have been considered. In that case also our proposed method (AADSFR50) achieves better accuracy.

Fig. 4 shows the sample of the output video that has been generated after implementing AADSFR50. Fig. 4(a) shows the true positive case where abnormal video was detected as abnormal. Fig. 4(b) shows the true negative case where normal video was detected as normal. Fig. 4(c) shows the false positive case where abnormal video was detected as normal. Fig. 4(d) shows the false negative case where normal video was detected as abnormal.

TABLE I. Comparison of Different Existing Algorithms and our Proposed Algorithm (AADSFR50)

Sr. No	Algorithm	Accuracy (%)
1	C3D[16]	23
2	TCNN[17]	28.4
3	3D Resnet 34[18]	27.2
4	3D ConvNets[19]	45
5	Semi- supervised GAN[20]	40.9
6	VGG-16[21]	72.66
7	VGG-19[21]	71.66
8	FlowNet[22]	71.33
9	Our proposed Algorithm (AADSFR50)	75

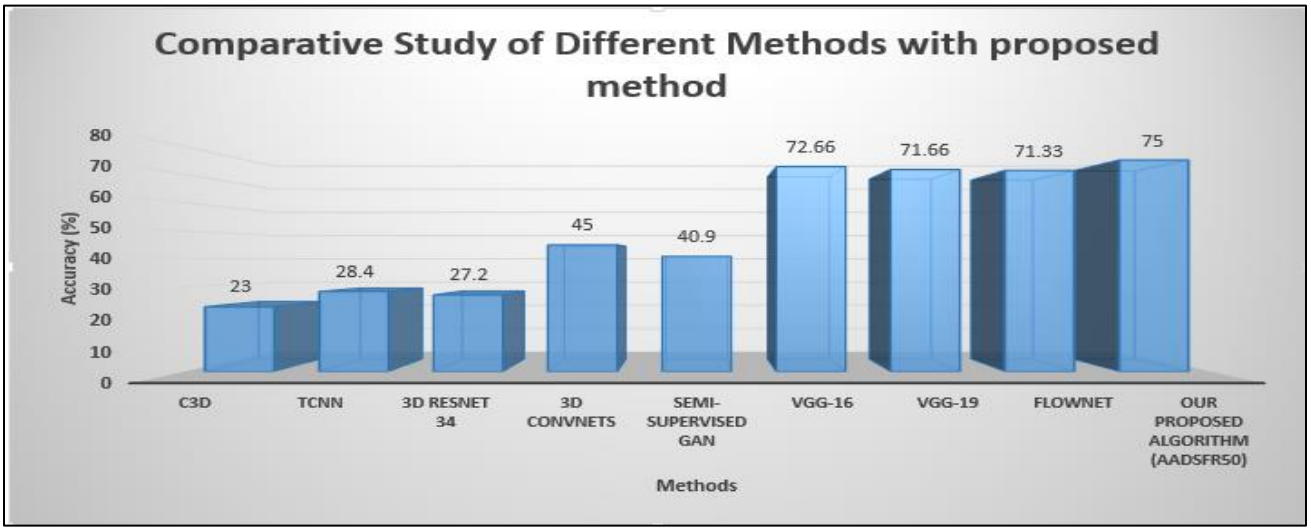
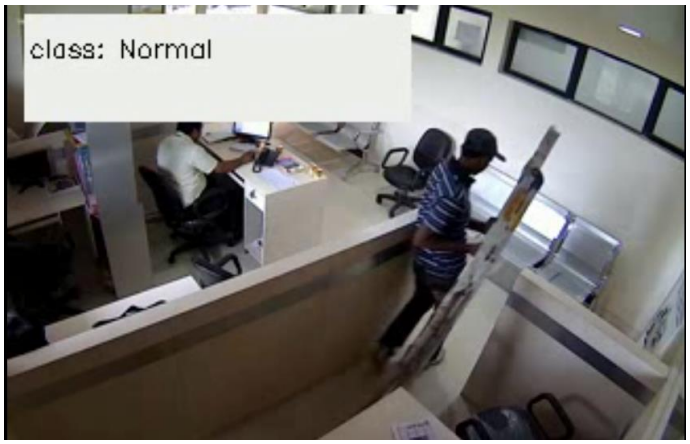


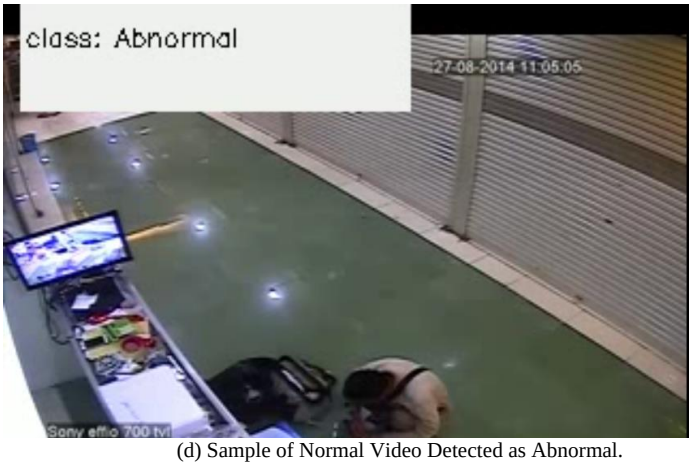
Fig. 3. Comparative Study of Different Methods with our Proposed Method (AADSFR50).

(a) Sample of Abnormal Video Detected as Abnormal.



(b) Sample of Normal Video Detected as Normal.

(c) Sample of Abnormal Video Detected as Normal.



(d) Sample of Normal Video Detected as Abnormal.

Fig. 4. Implementation Results of our Proposed Method (AADSFR50) - (a) Sample of Abnormal Video Detected as Abnormal (b) Sample of Normal Video Detected as Normal (c) Sample of Abnormal Video Detected as Normal (d) Sample of Normal Video Detected as Abnormal.

VI. Conclusion

Well known task has been found among researchers to detect abnormal activity from surveillance videos. Much progress has been achieved on this task but still, it is an open and interesting task for the researcher. Here in this paper, UCF crime dataset has been taken which contains challenging videos for the classification of abnormal and normal activities. SlowFast Resnet 50 has been used to perform the feature extraction. The fast pathway is used to analyze the static content and the Slow pathway is used to analyze the dynamic content of videos. Good accuracy has been achieved with the AADSFR50 method compared to state of art method and other existing approaches. For state of art method 47.8% accuracy increases and with other approaches 2.34% – 52% accuracy is increased. Still, there is the scope for applying a more optimized model such as Deeper Resnet to achieve more accurate results.

References

[1] Divya Thakur, Rajdeep Kaur, An Optimized CNN based Real World Anomaly Detection in Surveillance Videos, International Journal of Innovative Technology and Exploring Engineering (IJITEE) , vol. 8, pp.465-473, 2019.

[2] Guansong Pang, Chunhua Shen, Longbing Cao, Anton Van Den Hengel, Deep Learning for Anomaly Detection: A Review, ACM Computing Surveys,vol. 1, pp.1-36, 2020.

[3] Leman Akoglu, Hanghang Tong, and Danai Koutra, Graph based anomaly detection and description: a survey, Data Mining and Knowledge Discovery, vol. 29, pp. 626–688, 2015.

[4] Azzedine Boukerche, Lining Zheng, and Omar Alfandi, Outlier Detection: Methods, Models and Classifications, ACM Computing Surveys, vol. 53, pp. 1-37, 2021.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar, Anomaly detection: A survey, ACM Computing Surveys, vol. 41, pp. 1-72, 2009.

[6] Dr. Joy Iong Zong Chen, Dr. S. Smys, Social Multimedia Security and Suspicious Activity Detection in SDN using Hybrid Deep Learning Technique, Journal of Information Technology and Digital World, vol. 2, pp.108-115, 2020.

[7] Riddhi Sonkar, Sadhana Rathod, Renuka Jadhav, Deepali Patil, Crowd abnormal behaviour detection using deep learning, in Proc. of International Conference on Automation, Computing and Communication, pp. 1-5, 2020.

[8] Sarita Chaudhary, Mohd Aamir Khan, Charul Bhatnagar, Multiple Anomalous Activity Detection in Videos, in Proc. of 6th International Conference on Smart Computing and Communications, pp. 336-345, 2017.

[9] Laukasz Kaminski, Pawel Gardzinski, Krzysztof Kowalak, Slawomir Mackowiak, Unsupervised Abnormal Crowd Activity Detection in Surveillance Systems, in Proc. of 23rd International Conference on Systems, Signals and Image Processing, pp. 1-4, 2016.

[10] Federico Landi, Cees G. M. Snoek, Rita Cucchiara, Anomaly Locality in Video Surveillance, arXiv preprint arXiv: 1901.10364, pp. 1-5 , 2019.

[11] Shikha Dubey, Abhijeet Boragule, Moongu Jeon, 3D ResNet with Ranking Loss Function for Abnormal Activity Detection in Videos, in Proc. of International Conference on Control, Automation and Information Sciences, pp. 1-7, 2019.

[12] Allah Bux Sargano, Xiaofeng Wang, Plamen Angelov, Zulfiqar Habib, Human Action Recognition using Transfer Learning with Deep Representations, IEEE, in Proc. of International Joint Conference on Neural Networks, pp. 463-469, 2017.

[13] Waqas Sultani, Chen Chen, Mubarak Shah, Real-world Anomaly Detection in Surveillance Videos, Cornell University Library, arXiv:1801.04264, pp. 1-10, 2018.

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He, SlowFast Networks for Video Recognition, arXiv:1812.03982v3, pp. 1-10, 2019.

[15] The towardsdatascience website, Available: https://towardsdatascience.com/slowfast-explained-dual-mode-cnn-for-video-understanding, 2018.

[16] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., Learning Spatiotemporal Features with 3D Convolutional Networks, in Proc. of the IEEE International Conference on Computer Vision (ICCV), pp. 1-16, 2015.

[17] Hou, R.; Chen, C.; Shah, M., Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos, in Proc. Of the IEEE International Conference on Computer Vision (ICCV), pp. 5823–5832, 2017.

[18] Hara, K., Kataoka, H., Satoh, Y., Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition, in Proc. of the IEEE International Conference on Computer Vision Workshop (ICCV), pp. 3154–3160, 2017.

[19] Ramna Maqsood, Usama Ijaz Bajwa, Gulsha Saleem, Rana Hammad Raza, Anomaly Recognition from Surveillance Videos using 3D Convolution Neural Network, Multimedia Tools and Applications, vol. 80, pp. 18693-18716, 2021.

[20] Juan Montenegro and Yeojin Chung, Semi-supervised generative adversarial networks for anomaly detection, Innovative Economic Symposium 2021 – New Trends in Business and Corporate Finance in COVID-19 Era , vol. 132, pp. 1-8, 2022.

[21] Simonyan, K.; Zisserman, A., Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556.61., pp. 1-14, 2014.

[22] Ilg, E. Mayer, N., Saikia T., Keuper M., Dosovitskiy A., Brox, T., Flownet 2.0: Evolution of optical flow estimation with deep networks, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462–2470, 2017.