Anomaly Detection from Video Surveillances using Adaptive Convolutional Neural Network

Deepak Mane[1,1], Prashant Kumbharkar[2,] Poonam Pawar[3], Karishma Katkar[4],
Siddhali Shah[5], Khushi Jamwal[6]

[1,2,3,4,5,6] JSPM's Rajarshi Shahu College of Engineering, Pune-411033, Maharashtra, India
[1*]dtmane@gmail.com, [2]pbk.rscoe@gmail.com, [3]pam2pawar@gmail.com,
[4]kamma.katkar@gmail.com, [5]siddhalishah185@gmail.com, [5]khushijamwal173@gmail.com.

Abstract. Anomaly detection is finding various anomalous activities taking place in the video. Using an unsupervised learning technique, surveillance videos identify various real-time video anomalies in the dataset. In this paper proposed an adaptive convolutional neural network (ACNN) model of deep learning and classified the anomalies into two classes anomaly and non-anomaly. The two main key points that improve the learning algorithm of ACNN are: first, extract the hidden key features from the frames obtained by converting the videos to images using an adapted Convolutional Neural Network. Second, along with traditional CNN layers proposed model contains customized dropout and dense layers, which improves the performance during training. Performance evaluation is done through the UCF standard dataset. It consists of full-framed real-world videos obtained through CCTVs which are real anomalies like accidents on the roads, fire explosions, and fighting. Here, we predicted various anomalous activities in video anomaly detection by utilizing normal videos and videos containing abnormal activities. The proposed ACNN model produced better accuracy of 91% as compared to other existing algorithms used for the same dataset.

Keywords: Convolutional Neural Network, Pattern classification, deep learning, Anomaly detection.

1 Introduction

Anomaly detection in video frames is essential for surveillance systems due to its abnormal activities that rarely occur in real-time videos. The detection of anomalous illegal activities is an essential aspect of video surveillance. The main aim is to detect abnormal activities such as finding multiple moving entities with significantly less prior knowledge for video surveillance or detecting specific dangerous incidents such as fighting on the roadside captured in the CCTVs, explosions which cause fire, accidents on the roads, etc. The detection process focuses on detecting the video frames that consist of various anomalous activities among videos that the frames indicate. In contrast, segmentation focuses on determining which video frame is

[1] Deepak Mane
*dtmane@gmail.com

showing anomalous activity and which part of the frame is showing anomalous behavior [3]. Existing research and paper studies have focused mainly on finding the anomalies in traffic captured in the surveillance videos. In contrast, we can apply this method to detect many abnormal real-time situations happening in public places. Many researchers used variations of traditional CNN for different applications, which produce state-of-art results [16][17]. The applications of our proposed models are in monitoring the human activities in public places like railway and bus stations, other transportation facility areas like airports, multiplex, big shopping malls, roads, parking areas in public and private places, etc.

The aim of this paper is

- To spot and detect anomalies, including human-caused anomalous activities using the Adapted CNN.
- To recognize anomaly detection from each video this does not restrict n-dimensional data.

An overview of the paper is as follows: section 2 explains the literature survey of various related works in this field. Section 3 contains the proposed architecture, which explains the modules used and describes the ACNN algorithm. Experimental results are depicted in section 4, including model accuracy, model loss, confusion matrix, and classification report. Conclusion and future scope are explained in section 5.

2 Related Work

Anomaly detection comes under the unsupervised learning technique of machine learning which is meant for detecting the abnormal activities or patterns available in the dataset. In 2015, a method was proposed for detecting a real-time anomaly in crowded scenes. Experimental results prove that the algorithm is more time-efficient, and the system can immediately detect anomalies as soon as they occur in videos [1]. Extracting meaningful information from long videos is a challenging task solved in 2016, where two methods are built based on auto encoders first is to learn the features [2]. Automatically detecting abnormal events in long videos is a challenging task solved using generative models proposed in 2016. Also, this model uses convolutional extended short-term memory networks [3]. The most suitable model is chosen based on the prediction accuracy and reconstruction. CNN requires labels as learning signals, but the architecture presented in 2017 uses only two components, one for learning spatial features and the other for representing them [4]. In 2017 a model was proposed to solve the issue that occurs due to inverse problems in imaging by the classical iterative method [5]. Again in 2017, another method of anomaly detection using sparse coding was proposed. TSC is mapped with a stacked recurrent neural network that helps optimize the parameter and increases the speed of anomaly detection [6]. Though FNR works successfully in experimental results, significantly less information about its working mechanism is provided. However, this gap is filled in 2018 by building a connection between the Frobenius-norm-based and nuclear-norm-based representations [7]. A comparative analysis of deep learning-based anomaly detection methods was presented in 2018 [8]. The method proposed in 2017 introduces us to a deep CNN approach that captures the content of the whole image to study the correlation between the joints; immediate supervision of CNN is used[9]. After that, many modifications and innovations were made in anomaly detection, and noticible results were achieved.

To sum up, all the above papers aim to detect anomalies using various datasets. All the above research is done in various fields. Videos in the dataset are a crucial part of training as the accuracy depends on these factors. Therefore, the model proposed in this paper implements adaptive CNN on the UCF anomaly dataset, intending to detect video anomalies.

3 Proposed Architecture

In this paper, we proposed Adapted Convolutional neural network(ACNN) to train the dataset and to classify it into anomaly and normal events. The Fig 1 represents the system architecture of the ACNN model in a stepwise manner. The two main key points that improve the learning algorithm of ACNN are: first, extract the exact hidden key features from the frames obtained by converting the videos to images using an adapted Convolutional Neural Network. Second, along with traditional CNN layers proposed model contains customized dropout and dense layers, which improves the performance during training. However, proposed model does not restrict any n-dimensional features and handle large amount of data efficiently.

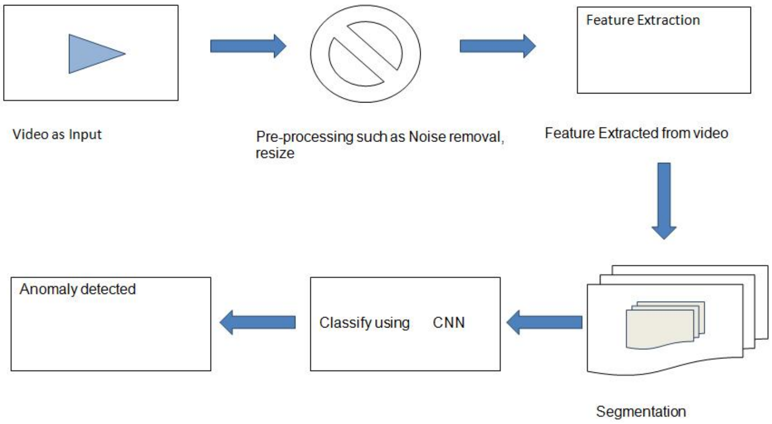

Fig. 1. System Architecture

Video input – video is passed as an input to the system, which is trained using CNN to detect anomalous activities.

Pre-processing- pre-processing involves applying various transformations on the raw data because the raw data, that is, the videos, may contain unnecessary noise disturbances, and some parts of videos maybe not be required, so clearing of these issues is done in pre-processing. The videos are a sequence of images and voice data; the video converted binary data is stored on the hard drive. Mathematically it can be expressed as,

Let D is the dataset. After pre-processing the dataset D’ (1)

Feature extraction- the videos are converted to images, and the images are converted to various frames. Feature extraction is the process of extracting or matching the feature present in one object to another object or fining it in that object. After learning the feature from a particular image frame, we try to find it in other image frames. We convert that extracted feature to numerical data and train the data based on these extracted features.

Segmentation- The videos are first segmented into images, and then the images are segmented into frames. These frames can extend up to thousands in numbers depending on the size of the videos. It is a process of breaking the images into sub-images called segments to decrease the complexity caused due to them to make further processing smooth. The 80 percent dataset is used for training, and 20 percent is used for testing. The videos of the training dataset are converted to frames and further passed to the algorithm.

Apply CNN- the training data is then passed to the Convolutional neural network. Proposed ACNN model does not restrict any n-dimensional features and handle large amount of data efficiently which is represented in Fig. 2.

3.1 ACNN learning Algorithm

Select the

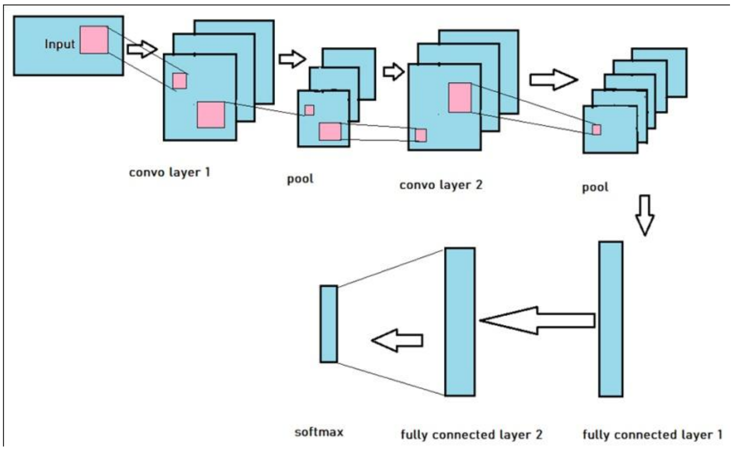

Fig. 2. ACNN architecture

3.1.1 Algorithm for ACNN

Forward steps of ACNN are explained using the following steps
Step I: Select the dataset containing anomalies.
Step II: Pre-processing the dataset, including allocating paths and dividing them into labels, image resizing, etc.
Step III: Splitting the dataset into training and testing data in proportions of 80% and 20%, respectively, to implement ACNN architecture.

Step IV: The output of the pre-processing step is given as an input to the customized convolutional layer, whose task is to extract different features present in the input images $C = I * f$ (2)

n equation (2), I is the input image, and f is the filter.

Step V: The output generated by the convolutional layer is passed to the pooling layer, which is used to decrease the rate of sampling by deducting the dimensions of the rectified feature map.

Step VI: images from earlier steps are passed to a fully connected layer where they are flattened

$$X = Z * V^T \quad (3)$$

Where in equation (3) X=row zero-mean data

Step VII: Dropout and dense are applied to prevent overfitting

4 Experiment Results

4.1 Dataset

The UMN dataset [10] has five differently arranged videos of people running and walking in different directions after some time. The feature that is used to classify the anomalous activity in [10] is the running activity. In [11], the UCSD ped1 contains 70 videos, and ped 2 contains 28 CCTV videos recorded simultaneously. Avenue [12] contains 37 videos again recorded at the same place, but these are short videos with few unrealistic anomalies. BOSS dataset [13] contains the suspicious activities performed by actors, for instance, panic circumstances and misbehaving with a person suffering from the disease. All these mentioned datasets are small in terms of time span and quantity. Moreover, they cover very unrealistic and less variety of anomalies. For our model training, we are using UCF- anomaly dataset, which contains 1000 videos and contains various activities like fighting, explosion, accidents, etc., which are realistic, and these videos are lengthy. Comparison of datasets represented in Table 1.

Table 1. Comparison of Datasets

	Video count	frames	length
BOSS	12	4068	25min
UMN	5	1300	5min
Avenue 37	900		30min
UCSD PED170	205		5min
UCSD PED228	163		5min
UCF-anomaly	1000	6000	5min

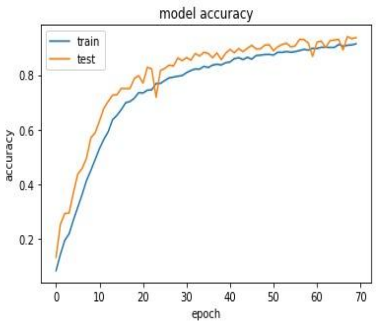

Fig 3. Model Training accuracy

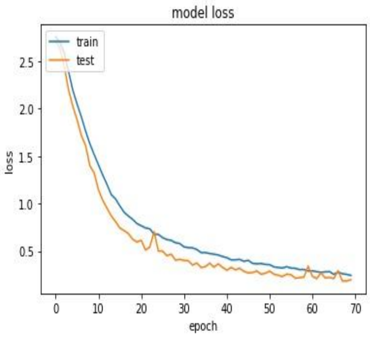

Fig 4. Model Training loss

4.2 Performance Evaluation

The final output layer of ACNN has two neurons which differentiate the entire dataset into two classes anomaly event detected and no event detected. The anomalies detected by our model are fighting, road accidents, and explosion. The videos are kept of original size without shortening them. We have trained our model on 80% of dataset and used 20% for testing our model. Each video generates more than 5000 frames. The model loss and model accuracy of the training and the testing data is as shown in the Fig 3 and Fig 4. To avoid over fitting further the dataset is manually trimmed and cleaned. To enhance the accuracy of the testing the training is increased. The mean squared distribution or mean squared error gives the average of squares of errors and the categorical cross entropy is the SoftMax function.

$$\text{MSE} = \frac{1}{n}\sum_{\boldsymbol{i}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i^2}) \tag{4}$$

$$\text{CCE} = \frac{1}{M}\sum_{p}^{M} -\log(\frac{e^{s_p}}{\sum_j^c e^{s_j}}) \tag{5}$$

The entire dataset is trained according to the training curve, and the training accuracy increases logarithmically. A confusion matrix is calculated for the actual and the predicted class, which evaluates our model's performance. The confusion matrix in Fig 5 shows that the accuracy for the anomaly class is 90%, and the non-anomaly class is 92%. The classification report is represented in Table 2. As per Table 3, ACNN has better accuracy than the existing methodology.

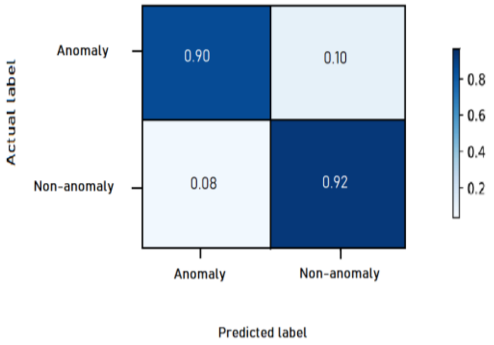

Fig 5. Confusion matrix

Table 2 Classification report.

Class	Precision	Recall	f-score	Support
Anomaly	90.23	91.58	90.54	25,584
Non-anomaly	91.45	90.02	90.96	36,897
Mean	90.84	90.08	90.55	∑= 62,481

Table 3. Result comparison with existing methodology

Reference	**Algorithm**	**Model Accuracy**
2008 [13]	Monitor based algorithm	75%
2015[18]	Un-supervised- RF, t(SNE)	85%
2018[19]	DONN+LSTM	89%
2019 [14]	LSTM convolutional autoencoder	87.30%
2020 [15	Two-stream convolutional networks model	80.47%
2021 [16]	ResNet and ConvLSTM	89%
Proposed model	**CNN algorithm**	**90.80%**

5. Conclusion

Due to the complex nature of these real-world anomalies, using only standard datasets may not be optimum for anomaly detection. This paper uses the ACNN model to detect the anomaly from real-time videos. Initially, we studied and analyzed the traditional techniques in deep learning, which detect real-world anomalous activities in surveillance videos. Here, we proposed an adaptive CNN model detect the suspicious activities observed in the CCTVs will assist in creating and maintaining the security in the public areas. The experimental results on the standard UCF dataset indicate that our proposed model for anomaly detection carries out more outstandingly than existing methods. The experiment results are proved by analyzing the test data results by calculating the confusion matrix and the classification report. Exclusively we have also mentioned the comparative table, which proves our proposed model is more accurate. In the future, we can reduce computational complexity using assembled deep learning techniques and use the best evaluation strategies.

References

1. Sabokrou, M., Fathy, M., Hosseini, M., & Klette, R.: Real-time anomaly detection and localization in crowded scenes. In. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 56--62, (2015).
2. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., & Davis, L.S.: Learning Temporal Regularity in Video Sequences. In. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733--742,(2016).
3. Medel, J.R., & Savakis, A.E.: Anomaly Detection in Video Using Predictive Convolutional Long Short-Term Memory Networks. ArXiv, abs/1612.00390. (2016)
4. Chong, Y.S., Tay, Y.H.: Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder. In: Cong, F., Leung, A., Wei, Q. (eds) Advances in

Neural Networks. Lecture Notes in Computer Science(), vol 10262, pp 189--196 Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59081-3_23

5. Diamond, S., Sitzmann, V., Heide, F., &Wetzstein, G. :Unrolled Optimization with Deep Priors. ArXiv, abs/1705.08041, (2017).
6. Luo, W., Liu, W., & Gao, S.: A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In. IEEE International Conference on Computer Vision (ICCV), pp. 341—349, (2017).
7. Peng, X., Lu, C., Yi, Z., & Tang, H.: Connections Between Nuclear-Norm and Frobenius-Norm-Based Representations. In. IEEE Transactions on Neural Networks and Learning Systems, 29, pp. 218—224, (2018).
8. Khan, Naimat& Wan, Wanggen.: A Review of Human Pose Estimation from Single Image. In. International Conference on Audio, Language and Image Processing (ICALIP),pp.230--236,10.1109/ICALIP.2018.8455796, (2018).
9. Ai, B., Zhou, Y., Yu, Y., & Du, S.: Human Pose Estimation Using Deep Structure Guided Learning. *In.* IEEE Winter Conference on Applications of Computer Vision (WACV*)*,pp. 1224—1231, (2017*)*.
10. Unusual crowd activity dataset of university of minnesota. http://mha.cs.umn.edu/
11. Lu, C., Shi, J., & Jia, J.: Abnormal Event Detection at 150 FPS in MATLAB. *In.*IEEE International Conference on Computer Vision, pp.2720-2727, (2013).
12. Wang, J., & Xia, L.:Abnormal behavior detection in videos using deep learning. Cluster Computing, pp. 1—11(2018)
13. Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D.: Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. In. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30,pp.555—560, *(2008)*.
14. Lu, Y., Mahesh Kumar, K., Nabavi, S.S., & Wang, Y.: Future Frame Prediction Using Convolutional VRNN for Anomaly Detection. In. 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1—8, (2019).
15. Hao, W., Zhang, R., Li, S., Li, J., Li, F., Zhao, S., & Zhang, W.: Anomaly Event Detection in Security Surveillance Using Two-Stream Based Model. In. Secur. Commun. Networks, pp. 8876056:1--8876056:15, (2020*)*.
16. Vosta, S., & Yow, K.C.: A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras. In. Applied Sciences, (2021).
17. Mane, D.T., & Kulkarni, U.V.: Visualizing and Understanding Customized Convolutional Neural Network for Recognition of Handwritten Marathi Numerals. In. Procedia Computer Science, 132, pp. 1123--1137, (2018).
18. Lundström, J., Morais, W.O., & Cooney, M.D.: A holistic smart home demonstrator for anomaly detection and response. In. IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 330—335, (2015).
19. Naseer, S., Saleem, Y., Khalid, S., Bashir, M.K., Han, J., Iqbal, M.M., & Han, K.: Enhanced Network Anomaly Detection Based on Deep Neural Networks. In. IEEE Access, 6, pp. 48231-48246, (2018).