Anomaly Detection Network for Video Surveillance Applications

R. Newlin Shebiah[1], J. Ajith[2], R. Hariharan[3]

Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi, 626 005, India.

---***---

Abstract - *Video surveillance system plays a vital role in the security and protection system of modern cities, since smart monitoring surveillance cameras equipped with intelligent video analytics techniques can monitor and pre-alarm for abnormal events. However, with the expansion of the video surveillance network, massive surveillance video data poses enormous challenges to the analytics, storage and retrieval in the big data. This paper presents an intelligent processing and utilization solution to the big video data based on the abnormal event detection and alarming messages for abnormal event detection from front-end smart cameras. The proposed methodology for anomaly detection includes motion fusion block aims at fusing motion information across frames, feature extraction by transfer learning and LSTM network. Experimental results reveal that our proposed approach can reliably pre-alarm security for abnormal events, substantially reduce storage space of recorded video and significantly speed up retrieval for evidence video.*

Key Words: AlexNet, Anomaly detection, Surveillance camera and LSTM

1. INTRODUCTION

Surveillance cameras are utilized out in the open spots for example boulevards, crossing points, banks, shopping centers, and so on to expand open security. The observing capacity of law implementation offices has not kept pace. Result, there is a glaring inadequacy in the usage of observation cameras and an unworkable proportion of cameras to human monitors. One of the essential venture in video surveillance is detecting abnormal events such as unlawful or crime activities. Normally, abnormal activities hardly ever arise as compared to normal activities. Consequently, to relieve the waste of time, growing clever computer vision and prescient algorithms for automated video anomaly detection is a want. The primary aim of a anomaly detection system is to timely signal an undertaking that deviates normal patterns and perceive the time of the happening anomaly. Consequently, anomaly detection may be considered as coarse level video expertise, which filters out abnormal events from normal patterns. An anomaly event is detected, it may in addition be categorized into one of the unique activities the usage of classification techniques. Some steps toward addressing anomaly detection is to broaden algorithms to detect a specific abnormal event, for example accident and violence detector. But, it's obviously that such solutions can't be generalized to detect different abnormal events. Real-time abnormal activities are complicated and numerous. It is tough to listing all of the possible abnormal events. It is desirable that the anomaly detection algorithm does not rely on any prior information about the abnormal events. Anomaly detection should be done with minimum supervision.

In [1] proposed "An anomaly detection Network for Video Surveillance" that consists of a neural network for anomaly detection with the aid of deeply attaining feature learning, dictionary learning in 3 joint neural processing blocks and sparse representation. The recurrent neural network to research the dictionary by way of adaptive iterative hard thresholding algorithm and sparse representation. In [2] proposed "A abnormal event detection in surveillance video: a compressed domain approach for HEVC". They works includes the compression format any provide useful information to solve the challenge compared to raw pixels. The advantage of the compression format is that it already contains some valuable clues for video analysis. In [4] proposed "Scene image classification method based on the Alex-Net model". This paper presents the algorithm of scene classification, it which fully learning the deep characteristics of the images based on the Alex-Net model. They used the Alex-Net model learning scene image features and extract the last layer with 4096 neurons of the Alex-Net model. Then, they used the Lib-SVM training model for scene image classification and compared with classification method based on the regression model. In [5] proposed "Abnormal event detection at 150 FPS in MATLAB". They proposed an efficient sparse combination learning framework, based on inherent redundancy of video structures.

In [6] proposed "A smart monitoring cameras driven intelligent processing to big surveillance video data". They works includes the intelligent pre-alarming for abnormal events using correlation model table, Smart storage for surveillance video and rapid retrieval for evidence videos.

2. PROPOSED METHODOLOGY

Figure 1 shows the block diagram of the proposed method for anomaly detection in video surveillance applications. It has training phase and testing phase. In training phase, features from a collection of normal and abnormal activities are used to build a model that effectively discriminate normal and abnormal activity. In testing phase, real time videos are captured by the CCTV camera, and it is fed to intelligent processing system. With the pre-learned model, the input videos are classified as normal or abnormal events. If any abnormal activity is detected then an alarm triggered.

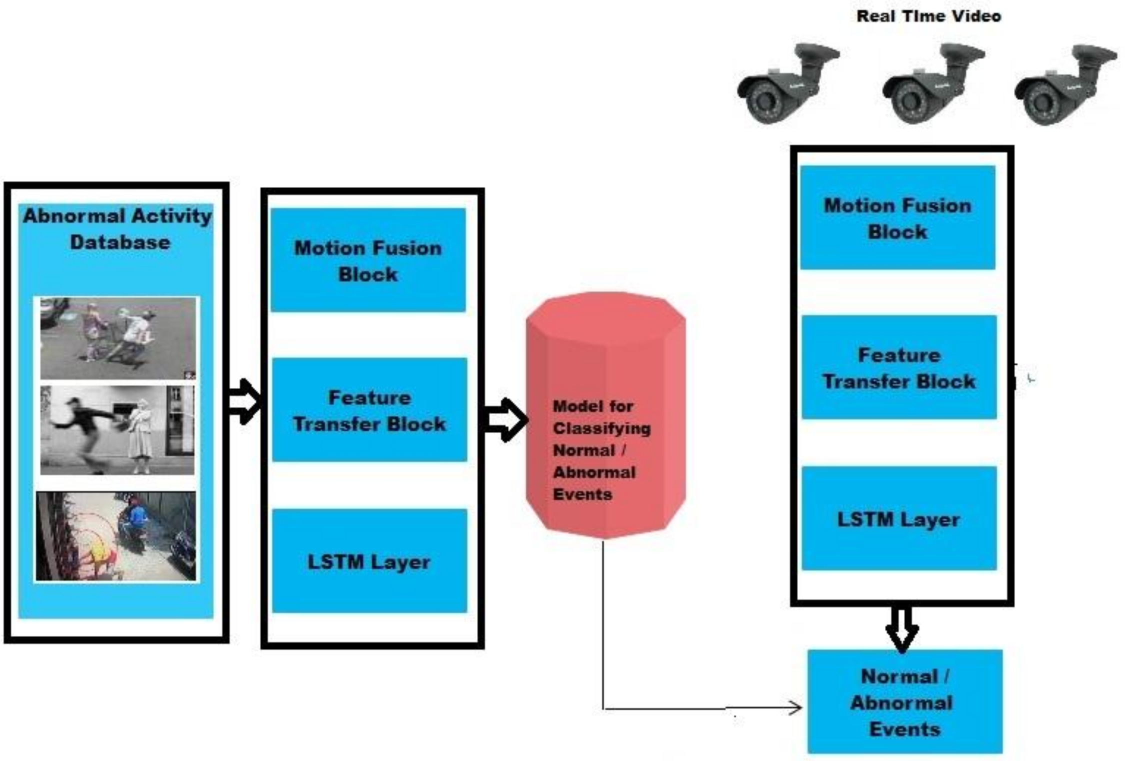

Fig. 1 Block Diagram of the Proposed Algorithm for Anomaly Detection

Motion Fusion Block

Motion Fusion blocks aims at fusion of motion fields across frames. The input video is converted in to frames and average between frames is concatenated to represent the motion fused image.

$$I'_t = \frac{(I_t + I_{t+1})}{2} \qquad (1)$$

$$I'_{t+2} = \frac{(I_{t+2} + I_{t+3})}{2} \qquad (2)$$

$$I'_{t+4} = \frac{(I_{t+4} + I_{t+5})}{2} \qquad (3)$$

The average of first two frames is considered to form the first plane. Average of third and fourth frame is considered as second plane and average of fifth and sixth plane is considered as third plane.

$$\text{Motion Fusion Block}_{\text{Average}} = [I'_t, I'_{t+2}, I'_{t+4}]$$

Motion fusion block can be considered as the difference between successive blocks as given by equation 4, equations 5 and equations 6.

$$d'_t = \frac{abs(I_t - I_{t+1})}{2} \tag{4}$$

$$d'_{t+2} = \frac{abs(I_{t+2} - I_{t+3})}{2} \tag{5}$$

$$d'_{t+4} = \frac{abs(I_{t+4} - I_{t+5})}{2} \tag{6}$$

$$\text{Motion Fusion Block}_{\text{Difference}} = [d'_t, d'_{t+2}, d'_{t+4}]$$

The motion fusion block fuses the temporal information across frames and effective in maintaining rich motion and appearance.

Feature Extraction Block

Discriminative features are extracted using transfer learning by Alexnet. Figure 2 shows the block diagram of Feature extraction using Transfer Learning.

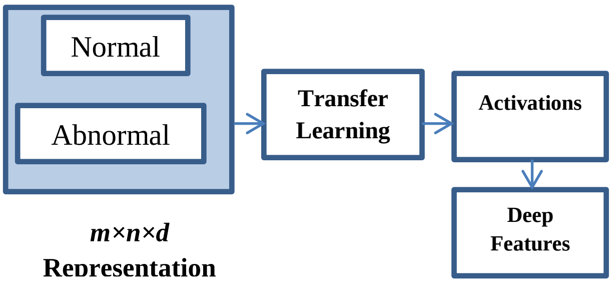

Fig. 2 Block Diagram of the Feature Extraction Module

The feature transfer block is to extract the features by using Alexnet architecture from the motion fusion block.

Alexnet Architecture

AlexNet is the first large scale convolutional neural network architecture. AlexNet architecture is a convolutional layer followed by pooling layer, normalization, convolutional-pooling-norm, and then a few more convolutional layers, a pooling layer, and then several fully connected layer. There are five convolutional layers, and two fully connected layers and finally fully connected layer going to the output classes.

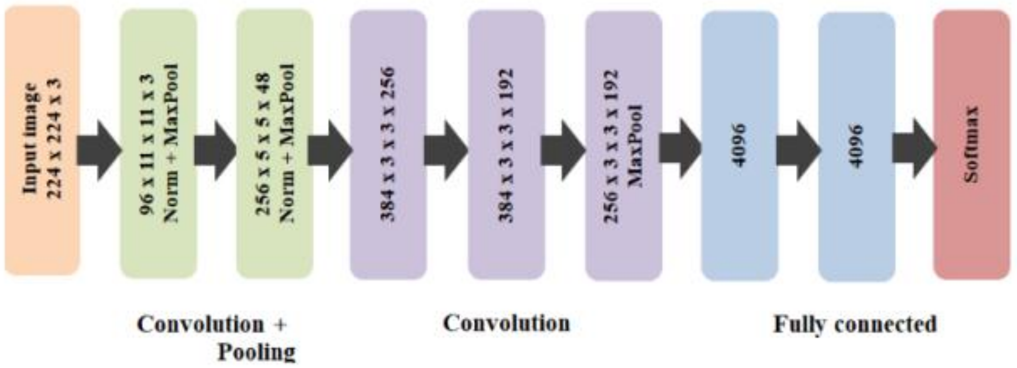

Fig. 3 Block Diagram of the AlexNet Architecture

AlexNet was trained on ImageNet, with inputs at size 227 x 227 x 3 images. If this first layer which is a convolutional layer for the AlexNet, it's 11 x 11 filters, 96 of these applied at stride 4. It's 55 x 55 x 96 in the output and 35K parameters in this first layer. The second layer is a pooling layer, it's 3 filters of 3 x 3 applied at stride 2. The output volume of the pooling layer is 27 x 27 x 96. The pooling layer does not learn anything because the parameters are trying to learn the weights. Convolutional layers have weights that they learn but pooling all they look at the pooling region, and they take the max. So there are no parameters that are learned. There are 11 x 11 filters at the beginning, five by five and some three by three filters. In the end, they have a couple of fully connected layers of size 4096. Finally, the last layer, FC8 is going to the softmax, which is going to the 1000 ImageNet classes. This architecture is the first use of the ReLu non-linearity.

LSTM Layer

LSTMs are designed to avoid the long-term dependency problem. All RNN (Recurrent Neural Networks) have the form of a chain of repeating modules of neural network. It is used for processing, predicting and classifying on the basis of time series data. Information is retained by the cells and the memory manipulations are done by the gates.

There are three gates. They are Forget gate, Input gate and Output gate. The forget gate is used by the information that no longer useful in the cell state is removed. The input gate is used to the Addition of useful information to the cell state. The Output gate is used the task of extracting useful information from the current cell state to be presented as an output. A LSTM network is a recurrent neural network. A recurrent neural network is a neural network and it attempts to model time or sequence dependent behaviour. This is performed by feeding back the output of a neural network layer at time *t* to the input of the same network layer at time *t + 1*.

3. RESULTS AND DISSCUSIONS

UCSD Pedestrian dataset is acquired with a stationary camera hooked up in an elevation and pedestrian walkways. It includes two sets: Pedestrain1 contains 7,200 frames with 40 abnormal events and Pedestrain2 contains 2,010 frames with 12 abnormal events, respectively. Videos are from the pedestrian walkways, recorded with a static camera at 10 frames per second. All other objects like vehicles (cycle, car etc.,) except for pedestrians are considered as abnormal event. Sample frames with normal and abnormal events from UCSD Pedestrian dataset is shown in Figure 4.

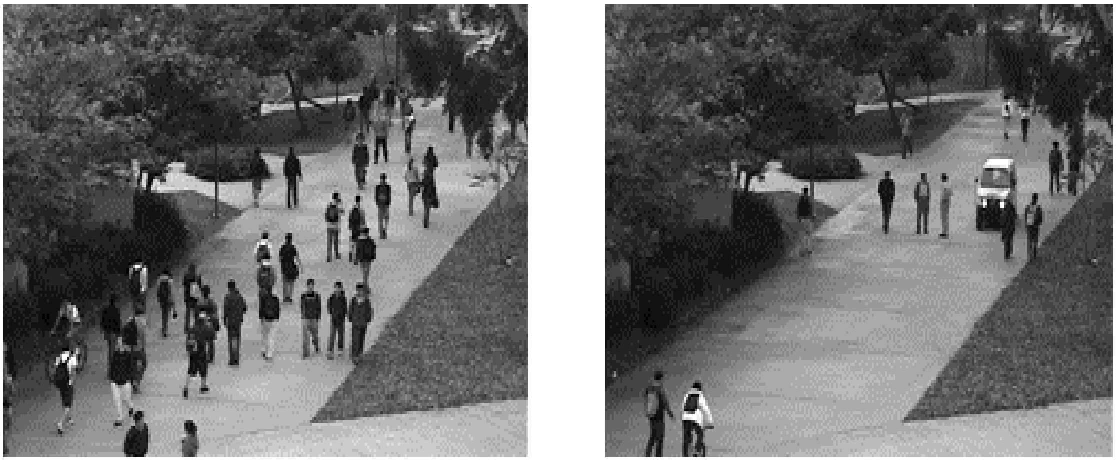

Fig. 4 Sample frames from UCSD Pedestrian dataset showing normal and abnormal activities

MOTION FUSION BLOCK OUTPUT

The frames from UCSD pedestrian dataset are processed in such a way to represent the motion information in a single image. Figure 5 shows the representation of motion fusion block.

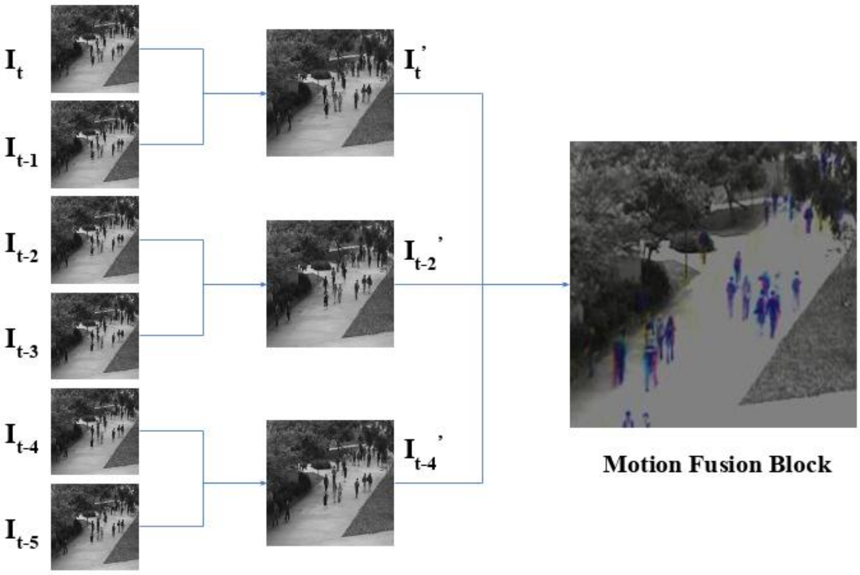

Fig. 5 Illustration of Motion Fusion Block

Figure 6 shows two adjacent frames and Figure 7 shows the average of frames concatenated to get an image in 3 planes and Figure 8 shows the differenced frames concatenated to form a single image.

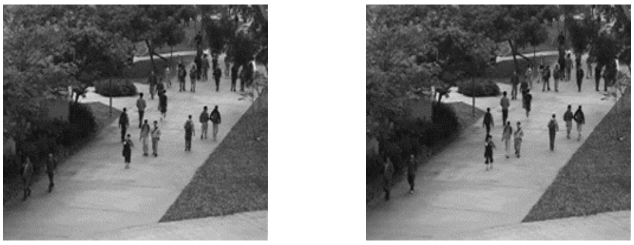

Fig. 6 Sample frames from UCSD Pedestrian at time *t* and *t+1*

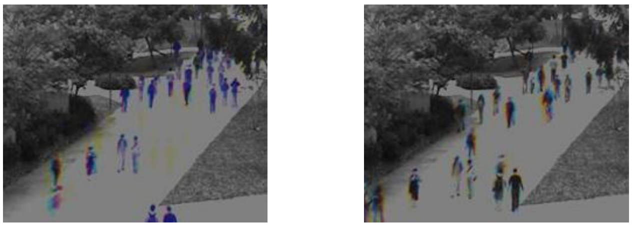

Fig. 7 Sample average image representation showing motion activities

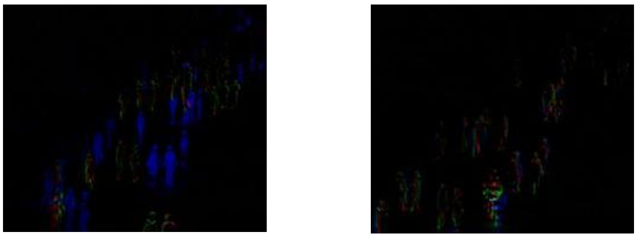

Fig. 8 Sample difference image representation showing motion activities

These images with dimension (*m×n×d*) are given as the input for transfer learning (Alexnet) for extracting the features. Then, the extracted features are given to the LSTM Layer.

The classification rate is calculated for normal and abnormal event images without using LSTM respectively. 70% of the data is used for training and 30% of the data is used for testing. Table 1 and Table 2 shows the confusion matrix of classifying normal and abnormal events using averaged representation and difference image representation. Here, the features are taken from Alexnet and classified. From the tables, it is inferred that the averaged image representation serves as a better representation compared with difference image representation.

Table 1 The Confusion matrix by using averaged images

	Abnormal	**Normal**	**Accuracy**
Abnormal	177	21	89.4%
Normal	25	419	94.4%
		Average	**92.8%**

Table 2 The Confusion matrix by using differenced images

	Abnormal	**Normal**	**Accuracy**
Abnormal	152	41	78.8%
Normal	50	399	88.9%
		Average	**85.8%**

Using features from LSTM layer recognition rate of 94.08% is obtained. Table 3 shows the recognition rate by varying activation function and optimizer. It is found that, the activations from FC6 with ADAM optimizer and batch size of 16 yield maximum accuracy.

Table 3 Recognition Rate by Varying Activation Function and Optimizer

Activations from	**Batch Size**	**Optimizer**	**Recognition Rate**
POOL5	16	ADAM	93.93%
POOL5	32	ADAM	93.93%
FC6	**16**	**ADAM**	**94.08%**
FC6	32	ADAM	92.68%
FC7	16	ADAM	80.53%
FC6	32	SGDM	76.95%

4. CONCLUSION

In this paper, a deep learning based framework for anomaly detection is proposed. Motion fusion block is designed to maintain the motion and appearance cues. The feature transfer block is used to extract the features by using CNN architectures and it exploiting the transferability of the neural network from different tasks/domains. The proposed model is effective in discrimination normal events from abnormal events.

REFERENCES

[1] Joey Tianyi Zhou , Jiawei Du , Hongyuan Zhu , Xi Peng , Yong Liu , Rick Siow Mong Goh , "AnomalyNet: An Anomaly Detection Network for Video Surveillance", Proc. IEEE Transactions on Information Forensics and Security .,pp. 2537-2550 , Feb 2019.

[2] Yihao Zhang, Hongyang Chao,"Abnormal Event Detection in Surveillance Video: A Compressed Domain Approach for HEVC", Proc . IEEE 2017 Data Compression Conf. (DCC).,pp. 2375-0359 , May 2017.

[3] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes", Comput. Vis. Image Understand., vol. 172, pp. 88–97, Jul. 2018.

[4] Jing Sun , Xibiao Cai , Fuming Sun, Jianguo Zhang, "Scene image classification method based on the Alex-Net model", 2016 3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS).

[5] C. Lu, J. Shi and J. Jia, "Abnormal event detection at 150 FPS in MATLAB", Proc. IEEE Int. Conf. Comput. Vis., pp. 2720-2727, Dec. 2013.

[6] Zhenfeng shao, jiajun cai, zhongyuan wang,"Smart monitoring cameras driven intelligent processing to big surveillance video data", IEEE Transactions on Big Data (Volume: 4 , Issue: 1, March 1 2018).

[7] Xingxing Zou, Jun Wen, "Detection of Object Security in Crowed environment", 2015 IEEE International Conference on Communication Problem Solving (ICCP).

[8] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury and L. S. Davis, "Learning temporal regularity in video sequences", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 733-742, Aug. 2016.