

Intensifying Security with Smart Video Surveillance

Swati Verma^{1,*}, Vikas Jaiswal¹, Radhey Shyam²

Abstract

The smart video surveillance system is designed to observe the activities of people nearby for safety concerns. This smart video surveillance can be installed at malls, public areas, banks, businesses, and ATM machines. The field of network surveillance research is rapidly expanding. The cause for this is the global insecurity that is occurring in great extent. As a result, an intelligent monitoring system that gathers data in real time, communicates, analyses, and interprets information about individuals being observed is required. Video surveillance systems are in great demand due to increasing safety concerns. The automated anomaly detection systems are much better since they require almost negligible human intervention and can more accurately detect anomalies. Besides this, the data can be easily stored and retrieved with the help of the video surveillance system. Video surveillance systems may be used for a variety of purposes, including traffic monitoring and anomaly detection. The study describes how three-dimensional deep neural networks are used to learn spatiotemporal elements of video feeds that can be used in our smart surveillance system.

Keywords: Automated anomaly detection, deep neural networks, spatiotemporal, video surveillance system, CNN

INTRODUCTION

Manual surveillance seems tedious and time consuming. The smart surveillance system can easily detect any irregular, inconsistent activity and that too without any human operator. The smart video surveillance system is a better approach than traditional surveillance systems or manual surveillance because they can detect any abnormal activities effectively for any crowded places, such as ATM machines, banks or any public areas which require constant monitoring. The anomalies detected by this system can be further used to increase security systems. Furthermore, these smart surveillance systems make anomaly detection more accurate and time efficient. In this study, an approach for both human anomaly detection and classification of single human activity recognition is proposed [1–5]. This approach can make the surveillance systems automated which means they will require less human intervention. We will propose a spatiotemporal auto encoder based on a 3D convolution network for this video surveillance project. The encoder collects spatial and temporal information, which is

subsequently used by the decoder to reassemble the frames. By estimating the reconstruction loss using the Euclidean distance between the original and reconstructed batches, aberrant occurrences are found. The underlying deep learning implementation technology used in the development of a smart surveillance system is discussed in this survey report [6–13].

*Author for Correspondence

Swati Verma

¹Student, Department of Computer Science and Engineering, Shri Ramswaroop Memorial College of Engineering and Management (SRMCEM), Lucknow, Uttar Pradesh, India

²Professor, Department of Computer Science and Engineering, Shri Ramswaroop Memorial College of Engineering and Management SRMCEM, Lucknow, Uttar Pradesh, India

Received Date: May 19, 2022

Accepted Date: June 04, 2022

Published Date: June 09, 2022

Citation: Swati Verma, Vikas Jaiswal, Radhey Shyam. Intensifying Security with Smart Video Surveillance. Recent Trends in Programming Languages. 2022; 9(1): 24–30p.

LITERATURE REVIEW

The study aims to put forward the previous works done and technologies used in the area of motion detection and image processing using deep learning techniques (See Figure 1).

The goal of an auto encoder-based anomaly detection in surveillance videos system is to detect anomalies in videos [4]. However, the system lacked few critical aspects. First of all, it makes an assumption that all the anomalous activities are already been covered in training dataset. This assumption limits the system as it will only cover few anomalous activities. Second problem of the system was that there is a deprivation of temporal continuity in test videos.

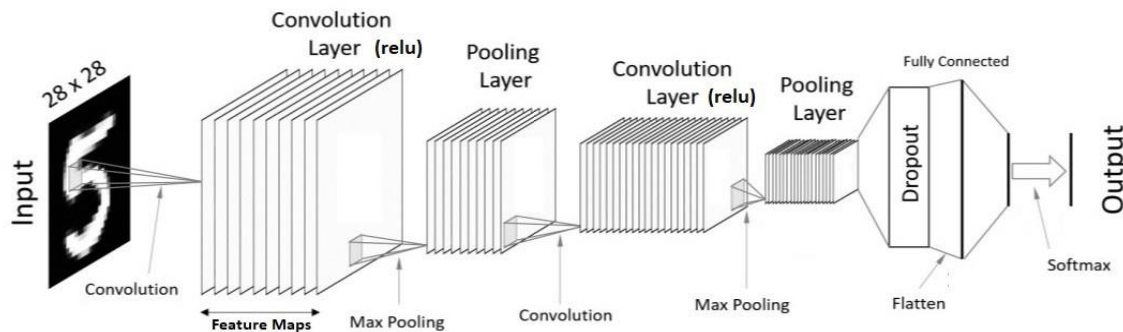


Figure 1. Typical architecture of a convolutional neural network (CNN) [11].

The study reviews about datasets, models, and proposed frameworks used in video surveillance system [5].

Spatiotemporal Auto Encoder

We used spatiotemporal auto encoder technique that uses deep neural network to learn video representation automatically and extracts features from both spatial and temporal dimension by performing 3D convolutions. The study proposes a system which recognizes anomaly using an encoder-decoder fully convolutional network integrating a spatiotemporal encoder [1].

3D Convolution Network

The paper discusses 3D CNN models that are used for action recognition [2]. The 3D feature extractors operate on spatial as well as on temporal dimensions and captures the motion information in the form of video streams. This CNN architecture generates several information channels from neighboring video frames, and each channel performs convolution and subsampling independently. By merging information from all channels, the final feature representation is generated. Similar approach has been used in our project where we perform convolution on each channel and combine the output obtained after convolution.

ConvLSTM

The paper mentions about features of LSTM and an advanced LSTM models with its additional features [3]. In each frame, we utilized LSTM to identify and forecast any unusual behavior (See Figure 2). ConvLSTM is a type of recurrent neural network that incorporates convolutional structures in both the input-to-state and state-to-state transitions for spatiotemporal prediction.

PROPOSED METHODOLOGY

The smart surveillance system acts in accordance with the following steps (See Figure 3):

1. The initial stage in this method is to extract frames from collected CCTV recordings. The work extracts the frame after a predetermined and short time period.
2. The pre-process input function is used to convert the resized picture to the format required by the model.
3. The auto encoder accepts high-dimensional input data and compresses it to a latent-space representation in the bottleneck hidden layer.
4. A normal transaction will be different from a fraudulent transaction. The auto encoder will have difficulty in reconstructing the fraudulent transaction, resulting in a large reconstruction error.

5. Multiple preprocessed frames are examined to provide the framework a feeling of the sequence. This section is then utilized to establish the final categorization.
6. Finally, after feature extraction it will identify if any anomalous activity is detected or not.

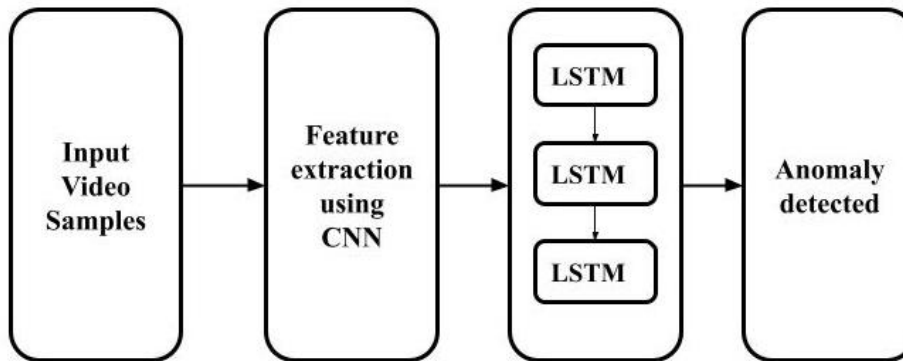


Figure 2. Working of LSTM.

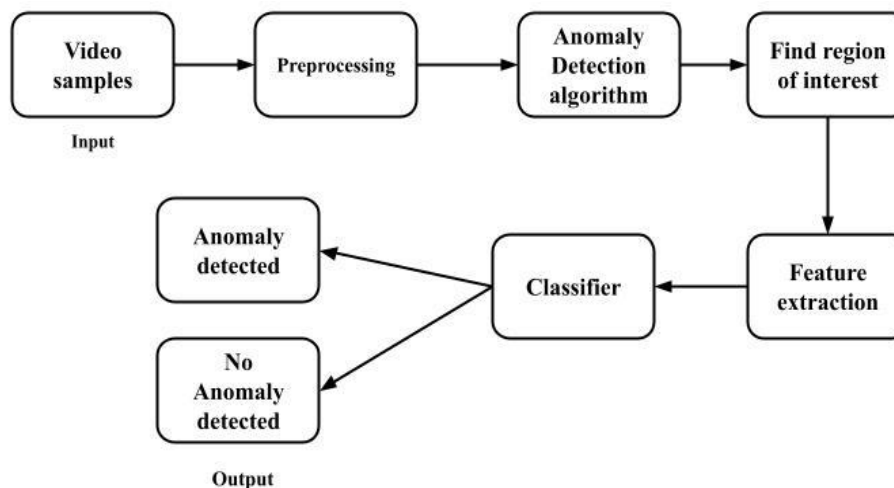


Figure 3. Workflow of the model.

IMPLEMENTATION

Datasets

The raw video data is preprocessed to improve the learning potential of the spatiotemporal auto encoder model. To reduce the dimensionality of the video data, it is first downloaded as a sequence of frames and then transformed to grayscale [6, 7]. The proposed model is trained using two of the most commonly used benchmarking datasets, UCSD Ped1 and UCSD Ped2 [8]. All of the training video sequences show just normal activity, whereas all of the testing video sequences show both normal and abnormal behavior. The strangeness of these two datasets is comprised of cyclists, skaters, carts, wheelchairs, and people wandering around on the grass. Each video clip in the Ped2 dataset has a variable number of frames.

UCSD Anomaly Detection Dataset

The UCSD Anomaly Detection Dataset was collected using a fixed camera at a height viewing pedestrian route. The number of persons on the paths ranged from a few to a large number as shown in Figures 4–7. The video only shows pedestrians in its basic setting. One of two causes might trigger unusual occurrences:

1. The movement of non-pedestrian creatures via pathways.
2. Pedestrian movement patterns that are out of the ordinary.

Peds 1: There is footage of people walking towards and away from the camera, as well as some perspective distortion. There are 34 training video examples and 36 testing video clips in this collection.

Peds 2: Scenarios with pedestrians moving in a straight line across the camera plane there are 16 video examples of training and 12 video samples of testing.



Figure 4. Peds 1: Cart.



Figure 5. Peds 1: Wheelchair.



Figure 6. Peds 1: Skater.



Figure 7. Peds 1: Biker.

Avenue Dataset for Abnormal Event Detection

The Avenue Dataset includes 16 training and 21 testing video clips. The videos were shot on the CUHK campus avenue and contain total of 30652 frames (15328 training, 15324 testing).

Normal scenarios are captured in the training videos. The testing movies depict both normal and aberrant situations. As follows are three aberrant samples in Figures 8–10.

Our dataset includes the problems listed below:

1. A slight camera shaking is seen.
2. There are a few outliers in the training data.
3. Normal patterns in training data are uncommon.

Preprocessing

The model turns incoming video data's raw frames into a suitable resolution. The pixel values are normalized to a range of 0 to 1 to ensure that all of the frames are on the same scale. To reduce processing complexity, the frames are also converted to grayscale [6]. Since deep learning models are data-driven, it is tough to get the right dataset for video anomaly detection. As a result, to increase the size of the training dataset in the temporal dimension, data augmentation is used [7]. This may be accomplished by combining many 10-frame strides from various train video sources in various sequences.

Spatiotemporal Auto Encoder

Auto encoders are a mix of encoders and decoders. It is an unsupervised artificial neural network that is used to learn efficient data coding. It works best when dimensionality reduction is required. In this case, the encoder outputs are less than the input.



Figure 8. Strange action.



Figure 9. Wrong action.



Figure 10. Abnormal object.

Three convolutional layers make up spatial encoders, whereas three deconvolutional layers make up spatial decoders. Because it retains the spatial relationships between pixels in a frame, convolution is preferred for learning spatial features. In CNNs, common hidden layers include convolutional layers, pooling layers, fully connected layers, and normalizing layers [9, 10]. In addition, a three-layer Convolutional Long Short Term Memory (ConvLSTM) model is used to create the temporal encoder-decoder [11]. A recurrent neural network captures the dynamic temporal behavior of time-sequence input data by using an internal memory to process the input sequences. Long short-term memory units are an upgrade of the RNN's generic building pieces 9. It is shown in Figure 11. When compared to the fully linked LSTM, convolutions replace matrix operations in ConvLSTM. When applied to the input-to-hidden and hidden-to-hidden connections, ConvLSTM produces superior spatial feature maps with fewer weights. ConvLSTM uses its convolutional layers to model spatiotemporal correlations [12].

To extract information from incoming image frames, video processing employs convolutional layers. By learning visual features from small squares of input data, the convolutional approach retains spatial connections between pixels. The dot product of the filter and the local region of the input picture frame is known as convolution in image processing. The disparity between the input and output representations is referred to as reconstruction error (error between input vector and output vector). The following formula calculates the reconstruction error of the recommended model [13]:

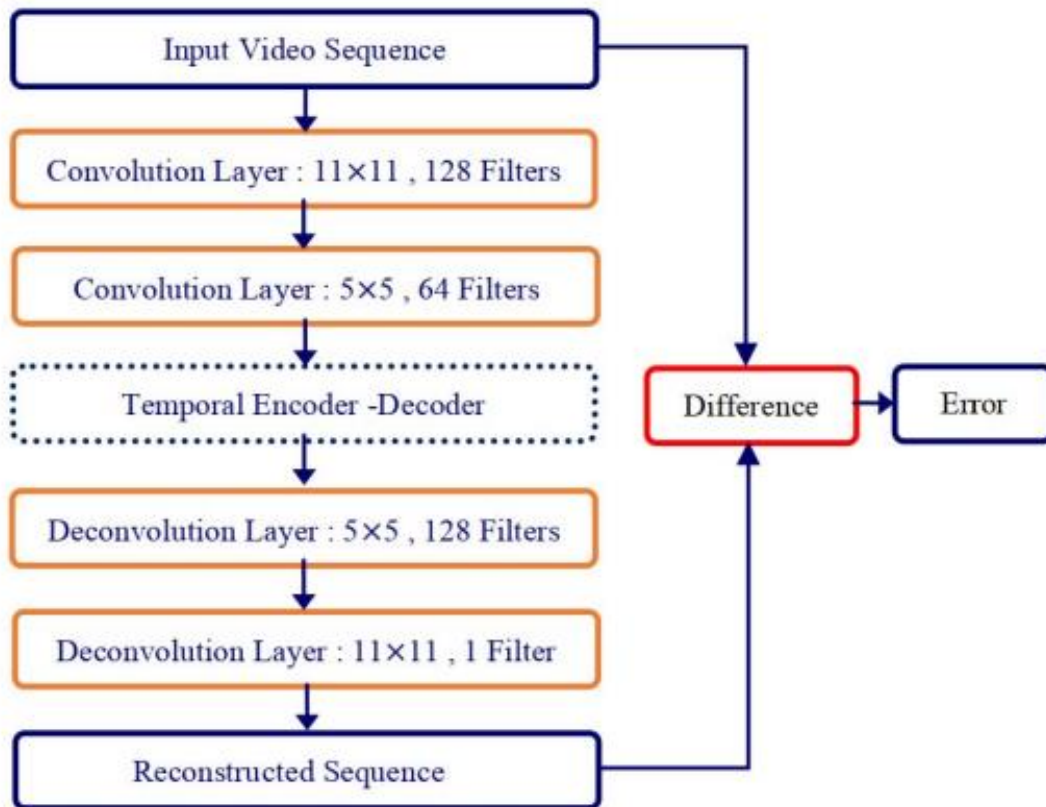


Figure 11. Architecture of proposed system.

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

Detection of an Anomaly

As previously stated, reconstruction error is employed to find abnormalities. It learns the reconstruction function for normal data, and we can use this model to detect anomalies. Normal data has minor reconstruction errors, whereas aberrant data has major reconstruction errors.

The threshold is used to establish if an event is anomalous or typical. It is a normal event if the reconstruction error is less than the optimal threshold; else, it is an abnormality. An anomalous event is detected when the reconstruction error exceeds the optimum threshold. When reconstruction error is less than the optimum threshold, then it is said to be a normal event.

ADVANTAGES

There are many advantages of this automated system, to name a few of them are:

1. When compared to traditional monitoring systems, smart video surveillance solutions need less time.
2. Data can be stored, accessed and retrieved more efficiently.
3. It is simple to set up and use, and it is more effective.
4. Provides a benefit of remote monitoring.
5. Requires almost no human intervention.

LIMITATIONS

There are numerous limitations and open challenges for video surveillance researchers. The problem with motion detection in dynamic scenes is that it is difficult to deal with changes in lighting and

weather, as well as shadow detection. A fast and accurate method is still needed to apply segmentation techniques to improve process performance. Monitoring many people or groups of people is difficult due to poor lighting, crowded environments, noisy images, and the presence of camera movements. Furthermore, frame level and pixel level detection are also a difficult task.

CONCLUSION

Smart surveillance system can serve in both public as well as in private organizations to maintain their security. The automated system is more reliable and time saving. Although many research and systems have been developed over the past few years, none of them has proved itself as reliable system as our smart surveillance system. Our smart surveillance system incorporates the expansion of deep neural networks to three dimensions in order to learn spatiotemporal aspects of the video feed. This approach makes our system better than the previous systems. However, few problems still exist in our system like with motion detection in dynamic scenes, to deal with illumination and weather changes, noisy images and shadow detection. It also assumes that all aberrant actions have previously been covered in the training dataset. This assumption restricts the system since it only covers a few abnormal actions. Another issue with the technology was a lack of temporal continuity in test footage.

REFERENCES

1. Du Z, Wu S, Huang D, Li W, Wang Y. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. *IEEE Trans Affect Comput.* 2019 Sep 10; 12(3): 565–578.
2. Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. *IEEE Trans Pattern Anal Mach Intell.* 2012 Mar 6; 35(1): 221–31.
3. Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on computer vision and pattern recognition.* 2018; 6479–6488.
4. Sahil Digikar, Abhijit Chaudhari, Pratik Angre, Rajat Pathak. Autoencoder Based Anomaly Detection in Surveillance Videos. *Open Access International Journal of Science & Engineering (OAIJSE).* 2021; 6: 29–32.
5. Shidik GF, Noersasongko E, Nugraha A, Andono PN, Jumanto J, Kusuma EJ. A systematic review of intelligence video surveillance: Trends, techniques, frameworks, and datasets. *IEEE Access.* 2019 Nov 25; 7: 170457–73.
6. Chong YS, Tay YH. Abnormal event detection in videos using spatiotemporal autoencoder. In *International symposium on neural networks*; Springer, Cham. 2017 Jun 21; 189–196.
7. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016; 733–742.
8. Mahadevan V, Li W, Bhalodia V, Vasconcelos N. Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition.* 2010 Jun 13; 1975–1981.
9. Radhey Shyam. Convolutional Neural Network and its Architecture. *Journal of Computer Technology and Applications (JoCTA).* 2021; 12(2): 6–14.
10. Vartika Srivastava and Radhey Shyam, Enhanced Object Detection with Deep Convolutional Neural Network, *International Journal of All Research Education and Scientific Methods*, 2021; 9(7): 27-36p.
11. Patraucean V, Handa A, Cipolla R. Spatio-temporal video autoencoder with differentiable memory. *arXiv preprint arXiv:1511.06309.* 2015 Nov 19.
12. Kiran BR, Thomas DM, Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *J Imaging.* 2018 Feb; 4(2): 36.
13. Radhey Shyam. Comparative Analysis of Distance Metrics using Face Recognition. *Research & Review: Discrete Mathematical Structures (RRDMS).* 2021; 8(1): 1–7.