International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

PREDICTION OF ANOMALOUS ACTIVITIES IN A VIDEO

Lekshmy K Nair

[1] *P.G. Student, Department of Computer Science and Engineering, LBS Institute of Technology for Women, Trivandrum, Kerala, India*

Abstract - *Ability to recognize and track abnormal human activities is one of the key challenges in video surveillance. Surveillance videos captures huge variety of realistic anomalies. With the huge amount of data available in the form of such video, it is impractical to manually analyze the behaviors of all the observed objects to quickly and correctly detect unusual patterns. In this paper, an intelligent video surveillance system capable of detecting anomalous activities within a video is proposed. Deep learning models are leveraged for this purpose. They can learn anomalies with the aid of training video dataset which contain both normal and anomalous events. Our proposed method uses CNN and a bidirectional LSTM for anomalous action recognition from videos. Convolutional neural network (CNN) has proved its effectiveness in applications such as action recognition, object detection, person detection etc. Recurrent neural network (RNN) and long short-term memory (LSTM) are widely used for learning temporal features from sequential data. The spatial features from frames in a video are first learnt by the CNN. Next, the temporal information is learnt by feeding the learnt features into a bidirectional LSTM network. Experiments are done on UCF Crime data using the proposed method and results have shown significant improvement in action recognition compared with the state-of-the-art approaches.*

Key Words: **Closed Circuit TV, Surveillance, Anomaly Detection, Video Analytics, Convolutional Neural Network, LSTM, Deep Learning.**

1. INTRODUCTION

The process of monitoring activities, behavior etc. for managing, influencing or protecting people is called Surveillance. A popular way of doing this is by observing the objects of interests from a distance with the help of electronic equipment's; for example, closed-circuit television (CCTV) cameras. CCTV cameras are widely used by organizations to help with safety and security. The analysis of data captured from CCTV cameras help organizations in preventing crimes or unexpected events before they occur.

1.1 Motivation

CCTV cameras produce a huge amount of video data. Effective surveillance can reveal anomalous events in these videos. Detecting crime occurrences, explosions, accidents etc. are some of the critical tasks in video surveillance. It is challenging to manually find anomalies from the huge amount of data available for surveillance purposes because of many reasons. Occurrence of anomalous events that are of interest for monitoring purpose would be very low compared to normal events. Hence manually analyzing this data to find out anomalous events which has very low probability of occurrence is a very tedious task that causes a considerable amount of man power wastage. Also, manual analysis can lead to human-generated errors.

1.2 Scope

In this paper, an intelligent video surveillance system is proposed which would help in such scenarios. Such a system can automatically analyze a video to detect anomalous activities. To make a system to detect anomalous activities, first it must be trained with anomalous activities. A major challenge in this method is the availability of the appropriate dataset. Anomalies are very contextual in nature and it also varies from domain to domain. So, we must use different datasets for different situations. Also, such a method will fall into the category of supervised learning which needs proper supervision and involved human efforts.

To deal with this, the proposed system is trained with normal activities with no or less supervision. This is not a big challenge as most of the activities in a surveillance video are normal. During testing. any activity which deviates from normal activity will be detected as abnormal or anomalous. Such a system will be domain free and can easily be used in any different scenarios.

This paper presents an unsupervised algorithm in which a deep learning model is trained with normal behaviors. Convolutional Neural Networks and bidirectional LSTMs are used for learning spatial and temporal features from training videos respectively. Once trained, it can be used to evaluate videos having normal and anomalous events. During testing, the model learns features from the test video and those with features deviating from normal are detected anomalous.

2. RELATED WORK

Anomaly detection is one of the most challenging problems in computer vision [6, 7, 11, 10, 9, 12, 13, 1]. For video surveillance applications, several researches were done to detect violence or aggression [17, 2, 3] in videos. [11] proposed a method that uses motion and limb orientation of people to detect human violence. Kooij et al. [2] employed video and audio data to detect aggressive actions in surveillance videos. Gao et al. proposed violent flow

descriptors to detect violence in crowd videos. More recently, Mohammadi et al. [3] proposed a new behaviour heuristic-based approach to classify violent and non-violent videos. Authors of [4] proposed a method in which the normal motion of people is used to create a model using tracking. Deviation from normal motion is then detected as an anomaly. But several of these approaches tend to avoid tracking action because of the difficulties in obtaining reliable tracks. Instead they learn global motion patterns through histogram-based methods [6], topic modelling [7], motion patterns [8], mixtures of dynamic textures model [10], Hidden Markov Model (HMM) on local spatio-temporal volumes [9], and context-driven method [11]. These approaches learn the patterns of normal actions from a training dataset that contains only normal events. Then the low probable patterns are detected as anomalies. Researchers in [12, 13] used sparse representation to learn the dictionary of normal behaviours after the success of sparse representation and dictionary learning approaches in several computer vision problems. During testing, the patterns which have large reconstruction errors are considered as anomalous behaviours. However, the sparse representation approaches tend to destroy the compositional information while grouping. Differentiating each feature into a pre-defined set of clusters is a tedious task. Also, these methods would require searching over a large space during testing time and makes it impractical for real world anomaly detection.

Deep learning has been found successful in areas of image classification, object recognition, object detection etc. Its ability to learn features have made it suitable for anomaly detection tasks. Deep learning model is an artificial neural network with several layers of hidden nodes. This model can be used for learning a hierarchical set of features from a given input data. There is no predefined set of features as in previous methods. Convolutional neural networks (CNN) have proved its effectiveness in a wide range of applications such as object recognition [9], person detection [12], and action recognition [10, 2]. A long short-term memory (LSTM) model is widely used for learning temporal patterns and predicting time series data [18]. CNN and LSTM together can learn spatial and temporal features in a video.

3. PROPOSED ALGORITHM

3.1 Description of the Proposed Algorithm

This section explains the proposed algorithm and its main components. Proposed algorithm consists of two major functions. First, the features in the frames of the input video is extracted by Convolutional Neural Network (CNN). Second, action in a set of frames is fed through bidirectional LSTM to study the temporal sequence. The output of the bidirectional LSTM module is analysed to recognise actions into normal/anomalous activity classes. The proposed algorithm consists of three main steps.

Step 1: Feature Learning

The features are extracted from the video with the help of a convolutional neural network (CNN). CNNs are widely used for object recognition and image classification tasks due to its ability of learning spatial features from an image. CNN is used to learn spatial features from each individual frame in the video. A CNN learns hidden patterns in images. So, it can capture even the smallest changes from frame to frame.

Training a deep learning model for image representation requires thousands of images and requires high processing power such as GPU for the weight adjustment of the CNN model. Getting the required model using this strategy is an expensive process, which is solved using a trained model can be used for other purposes. In the proposed algorithm, we used parameters of the pretrained CNN model, called VGGNet for feature extraction, which is trained on large scale ImageNet dataset of more than 15 million images. The architecture of the model is given in Figure 1. VGG is a convolutional neural network model proposed by K. Simonyan and A. Zisserman [19]. It became known as VGGNet.

On ImageNet dataset, VGGNet achieves a top-5 test accuracy of 92.7%. ImageNet consists of over 14 million images belonging to 1000 classes. The macro architecture of VGG16 can be seen in Figure 1. The network has 16 convolutional and fully connected layers. It performs 3x3 convolutions and 2x2 pooling from the beginning to the end. This extremely homogenous architecture is a specialty of VGGNet. The VGGNet consists of layers which perform 3x3 convolutions with stride 1 and pad 1. The POOL layers do 2x2 max pooling with stride 2 and no padding. The features of each frame are inputted to the next LSTM layer.

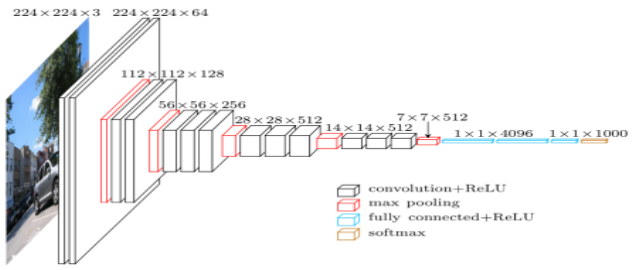

Fig – 1: Architecture of VGG16 Net

Step 2: Learning temporal features

Recurrent Neural Networks is used to find the hidden sequential or temporal sequences in the input data. Each input is considered independent in a traditional feedforward neural network. However, in case of video sequences, learning temporal dependencies between inputs are equally important as learning spatial features. RNNs can be used to learn such sequences. That means the output of an RNN is influenced by the history data as well in addition to the current input. Though RNNs can interpret sequences, in case

of long video sequences it tends to forget earlier inputs. This problem is called the vanishing gradient problem. For this, a special type of RNN, called LSTM [18] is used.

LSTMs are capable of learning temporal patterns from long sequences due to its special structure. It has input, output, and forget gates to help with the long-term sequence pattern identification. The gates are adjusted by a sigmoid unit that learns during training where it is to open and close. Eq. 1 to Eq. 7 [20] explain the operations performed in LSTM unit, where x_t is the input at time t. f_t is the forget gate at time t, which clears information from the memory cell when needed and keeps a record of the previous frame whose information needs to be cleared from the memory. The output gate o_t keeps information about the upcoming step, where g is the recurrent unit, having activation function "tanh" and is computed from the input of the current frame and state of the previous frame s_{t-1}. The hidden state of an RNN step is calculated through tanh activation and memory cell c_t. As the action recognition does not need the intermediate output of the LSTM, we made final decision by applying Softmax classifier on the final state of the RNN network.

$$i_t = \sigma((x_t + s_{t-1})W^i + b_i) \quad (1)$$

$$f_t = \sigma((x_t + s_{t-1})W^f + b_f) \quad (2)$$

$$o_t = \sigma((x_t + s_{t-1})W^o + b_o) \quad (3)$$

$$g = tanh((x_t + s_{t-1})W^g + b_g) \quad (4)$$

$$c_t = c_{t-1} \cdot f_t + g \cdot i_t \quad (5)$$

$$s_t = tanh(c_t) \cdot o_t \quad (6)$$

$$final\_state = soft\ max(Vs_t) \quad (7)$$

The features representing the video sequence of a time interval, T that is received from the previous step is fed to a bidirectional-LSTM. In bidirectional LSTM, the output at time t is not only dependent on the previous frames in the sequence, but also on the upcoming frames [21]. Bidirectional RNNs are quite simple, having two RNNs stacked on top of each other. One RNN goes in the forward direction and another one goes in the backward direction. The combined output is then computed based on the hidden state of both RNNs. Figure 2 shows the overall structure of bidirectional LSTM used in the proposed method. In normal time order, input sequence is fed into one network and in reverse time order for another. The outputs of the two networks are usually concatenated at each time step. This structure allows the networks to have both backward and forward information about the sequence at every time step which makes it more effective than other models which uses LSTM network for learning temporal action sequence.

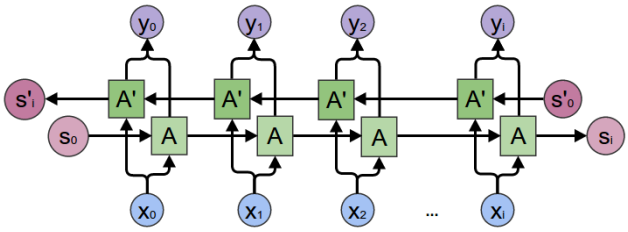

Fig – 2: Structure of a bidirectional LSTM

3.3 Training and Testing Phase

During training phase, the model is fed with only normal videos. Each video is then segmented into frames. The CNN first learns the spatial features from each of these frames. When it finishes processing a set of frames, say 20, it feeds the feature vector into the bidirectional LSTM network. The bidirectional LSTM network will then learn the temporal features from the set of frames. The learning will be done for several epochs (e.g. 1000 epochs) and by the end of the training the model will learn the feature vector for normal activities. Softmax classifier is used.

During testing phase, the model is fed with both normal and abnormal videos. It calculates the spatial and temporal features of the test videos like in testing phase and the output of bidirectional LSTM network is analyzed to detect if the test video falls into normal/anomalous activity class. The anomaly score value which differentiates a normal video from anomalous video can be found during the experimentation process.

4. RESULTS AND DISCUSSIONS

This section put forward details about the simulation results of the proposed method.

4.1 Dataset

Some of the popular datasets used for anomaly detection are briefly explained here. The UMN dataset [2] consists of five different staged videos where people walk around and after some time start running in different directions. The anomaly is characterized by only running action. UCSD Ped1 and Ped2 datasets [10] contain 70 and 28 surveillance videos, respectively. These videos were captured by a stationary camera mounted at an elevation. The camera overlooks pedestrian walkways. The normal videos in this dataset contains only pedestrians. Abnormal events are due to presence of non-pedestrian entities on walkways or anomalous pedestrian motion patterns. As we can infer, this dataset does not reflect realistic anomalies in video. Avenue dataset [12] consists of 16 training and 21 testing video clips. Although it contains more anomalies, they are performed by actors and captured at one location. Like [10], videos in this dataset are short and some of the anomalies are unrealistic (e.g. throwing paper). Subway Exit and

Subway Entrance datasets [22] contain one long surveillance video each. The two videos capture simple anomalies such as walking in the wrong direction and skipping payment. In general, the popular datasets for anomaly detection are limited in terms of the number of videos and length of the video. Also, most of them are staged and some anomalies are unrealistic. In this paper, the dataset used is the UCF-Crime dataset. It consists of long untrimmed surveillance videos which cover both normal activities and anomalous activities such as Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism [16]. These anomalies are selected because they have a significant impact on public safety.

Fig - 3: Sample anomalous actions from UCF Crime Dataset

This dataset is used for two tasks in the paper. First, the deep learning model is trained with normal activities available in the dataset. Then testing of the model is done using anomalous activities in the dataset. Sample images of anomalous frames in the UCF-Crime dataset is given in Figure3.

4.2 Simulation Results

In the section, the results obtained when the model is tested with a group of test videos is explained. The set consists of both normal and abnormal activity videos. For estimating test accuracy, 100 videos are taken. In this set 64 videos are videos containing anomalous activities and 36 are normal videos.

Table 1 shows experimental results using the proposed method. Out of the 64 abnormal videos, the proposed method classified 63 as abnormal and 1 as normal. Also, out of the 36 normal videos, 35 are classified as normal and 1 as abnormal. The abnormal videos in the test dataset contained anomalous activities such as Abuse, Shooting, road Accident, Explosion and Robbery. The Accuracy is calculated as (TP+TN)/total = (52+29)/100 = 81**%.** TP is the number of actual anomalous videos that are predicted as anomalous. TN is the number of actual normal videos which is predicted as normal. The erroneous predictions are due to the inadequacy of differences in features that distinguishes between normal and anomalous activity classes. The below table shows the confusion matrix of the experiments done on different abnormal activities in the UCF-Crime dataset.

N=100	Predicted: NO	Predicted: Yes	
Actual: NO	TN=29	FP= 7	36
Actual: YES	FN=12	TP=52	64
	41	59	

Table - 1: Confusion Matrix

The following bar graph representation shows the accuracy obtained for different action categories.

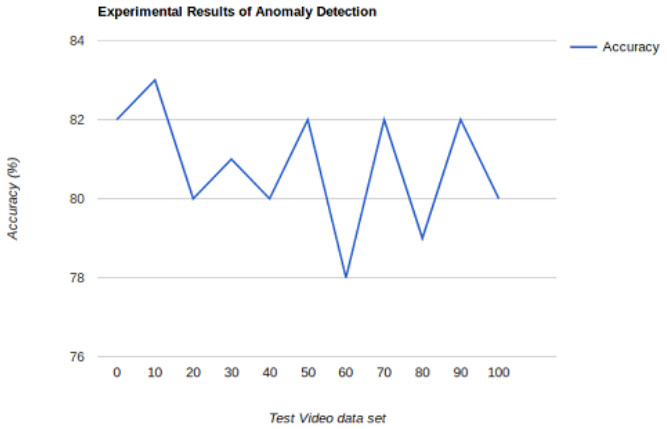

Chart - 1: Test Accuracy of the proposed method

4.3 Implementation Details

In the training phase, videos are first segmented into frames using opencv. Features are extracted from each frame with the help of VGG Net. The CNN takes one frame at a time as input. Once the number of frames denoted by the value MAX_ALLOWED_FRAMES has been processed, the encoded features of these frames are fed into bidirectional LSTM module for motion encoding. The spatial features are extracted from the fully connected (FC) layer FC3 of the VGGNet. The parameter BATCH_SIZE is for mini-batch gradient descent used for the LSTM training. The optimizer used is RMSProp optimizer. Root Mean Square Propagation (RMSProp) maintains per-parameter learning rates that are adapted based on the average of recent magnitudes of the gradients for the weight. Mini-batches of size 64 is used and each training volume is trained for a maximum of 20 epochs. Softmax is used as the activation function for the output layer and the categorical cross entropy as the loss function. The deep learning algorithms are implemented using Keras library with backend as Tensorflow.

Software Prerequisites- Python 3.x, Tensorflow, Keras, Numpy, h5py, pillow, patool, opencv-python, scikit-learn, matplotlib.

5. Comparison of Machine learning techniques for Human Activity classification

In this section, the proposed method is compared with few state-of-the-art approaches for anomaly detection. Dictionary based approaches are widely used in anomaly detection [12]. In this, normal videos are used in training phase to learn features and create and the reconstruction error at the time of testing would detect anomalies. Convolutional neural networks are well known for its image classification capabilities. Fully convolutional feed-forward deep auto-encoder based approaches are also used to learn local features and classify [1]. During training phase, the network is fed with normal videos and features are extracted. In the testing phase, the videos having events deviating from normal are expected to have high reconstruction error. Binary SVM classifier can also be used for classification tasks. Features are computed for each video, and a binary classifier is trained with linear kernel. The performance comparisons in terms of accuracy are shown in Table 2. The results show that the proposed approach significantly outperforms the existing methods.

Dictionary based approaches [12] is not robust enough to discriminate between normal and anomalous pattern. It sometimes produces low reconstruction error for abnormal videos. Even though the method proposed which uses Fully convolutional feed forward deep auto-encoder [1] learns normal patterns effectively, it tends to produce high anomaly scores even for new normal patterns. Proposed method produces high anomaly scores for the anomalous frames and hence is suited for anomaly detection tasks.

No.	Algorithm Used	Accuracy
1	Binary classifier	50.0
2	Fully convolutional feed forward deep auto-encoder [1]	50.6
3	Dictionary based approaches [12]	65.51
5	Proposed Algorithm	81

Table - 2: Accuracy comparison of various approaches for anomaly detection

6. CONCLUSIONS

The paper proposed a deep learning approach to detect real-world anomalies in surveillance videos. The deep learning algorithms such as CNN and bidirectional LSTMs are used to learn spatial and temporal features from UCF-Crime dataset. Most of the previous researches on video anomaly detection were done on unrealistic and staged video datasets. Because of that it may fail to give accurate output in case of realistic anomalies. The UCF-Crime dataset in contrast had untrimmed video sequences which contain both anomalous and normal events. This makes it more effective for video anomaly detection tasks. The advantage of the proposed method is that it is semi-supervised. Training data required is only of the normal events expected. Getting normal videos is not challenging as the occurrence of normal events is very much high compared to abnormal events. In future, a module can be further added to the proposed method which can classify the already classified anomalous activities into various activity classes. It will be a supervised module as the module must be trained with different labelled datasets so that it can identify activity.

REFERENCES

[1] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, June 2016.

[2] J. Kooij, M. Liem, J. Krijnders, T. Andringa, and D. Gavrila. Multi-modal human aggression detection. Computer Vision and Image Understanding, 2016.

[3] S. Mohammadi, A. Perina, H. Kiani, and M. Vittorio. Angry crowds: Detecting violent events in videos. In ECCV, 2016.

[4] S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In CVPR, 2010.

[5] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In CVPR, 2008.

[6] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. Abnormal detection using interaction energy potentials. In CVPR, 2011.

[7] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In ICCV, 2009.

[8] I. Saleemi, K. Shafique, and M. Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. TPAMI, 31(8):1472–1485, 2009.

[9] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In CVPR, 2009.

[10] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. TPAMI, 2014.

[11] Y. Zhu, I. M. Nayak, and A. K. Roy-Chowdhury. Context aware activity recognition and anomaly detection in video. In IEEE Journal of Selected Topics in Signal Processing, 2013.

[12] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.

[13] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In CVPR,2011.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.

[15] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. Learning deep representations of appearance and motion for anomalous event detection. In BMVC, 2015.

[16] Waqas Sultani, Chen Chen, Mubarak Shah. Real-world Anomaly Detection in Surveillance Videos. Center for Research in Computer Vision (CRCV), University of Central Florida (UCF).

[17] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu. Violence detection using oriented violent flows. Image and Vision Computing, 2016.

[18] K. Greff, R. K. Srivastava, J. Koutník, B.R. Steunebrink, and J. Schmidhuber, ‘‘LSTM: A search space odyssey,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[19] Very Deep Convolutional Networks for Large-Scale Image Recognition K. Simonyan, A. Zisserman ICLR 2015.

[20] H. Sak, A. Senior, and F. Beaufays, ‘‘Long short-term memory recurrent neural network architectures for large scale acoustic modeling,’’ in Proc. 15th Annu. Conf. Int. Speech Commun. Assoc., 2014, pp. 338–342.

[21] A.Ogawa and T.Hori,‘‘Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks,’’ Speech Commun., vol. 89, pp. 70–83, May 2017.

[22] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. TPAMI, 2008.