Suspicious Activity Detection through CCTV

[1]Prof. Akshay Agarwal, [2]Mr. Swapnil Galhate, [3]Ms. Shipra N. Suvarna
[1]Asst.Professor, [2,3]UG Student, [1,2,3]Department of Information Technology, Universal College of Engineering, Kaman, Maharashtra, India.
[1]akshay.agarwal@universal.edu.in, [2]swapnilgalhate1036@gmail.com, [3]987shipra@gmail.com

Abstract This report describes the look, implementation and evaluation of a system to detect suspicious activities. Activity detection has great importance in many applications particularly within the surveillance industry The detection of suspicious behaviors in crowded scenes deals with many challenges. Several specific solutions exist already for detecting suspicious activities. However, the need for these forms of systems are growing rapidly because it reduces human efforts. This project represents a system based on the identification of the human body and object detection and tracking. This paper presents an efficient method for detection and localization of unusual activities in videos. A pre-trained supervised FCN is modified into an unsupervised FCN with the help of fully convolutional neural networks (FCNs) and temporal data, which in turn detects the anomalies in scenes. By reducing computational complexities, accuracy and high performance speed is achieved. This was possible by carrying out a systematic formal inquiry and examining the cascaded detection. Experimental results suggest that detection and localization of the proposed method outperforms existing methods in terms of accuracy.

Keywords — Anomaly Detection, FCN.

I. INTRODUCTION

The use of surveillance cameras requires that computer vision technologies need to be involved for the analysis of very large volumes of video data. Unusual activity detection in captured scenes is one of the applications in this area. Localization and detection of anomaly is a challenging task in video analysis. This paper proposes and evaluates a different and new method for suspicious activity detection. Here we introduce and study a modified pre-trained convolutional neural network (CNN) for detecting and localizing anomalies. For processing a video frame [1] demarked a method in which the frames were first divided into a set of patches, then the anomaly detection was organized based on levels of patches. In difference to that, the input provided by the CNN algorithm is a full video frame in this paper. The new method is methodically simpler but faster in both the training and testing phase where the accuracy of suspicious activity detection is comparable to the accuracy of the method presented in [1]. The reference models include normal motion or shapes of every region of the training data. In the testing phase, those regions which differ from the normal model are considered to be abnormal. Classifying these regions into normal and abnormal requires extensive sets of training samples in order to describe the properties of each region efficiently. We have used Trajectory-based methods to define behavior's of objects. These trajectory-based methods have two main disadvantages. CNNs proved recently to be useful for defining effective data analysis techniques for various applications, but CNNs are computationally slow. .Due to these difficulties, there is a recent trend to optimize CNN-based algorithms in order to be applicable in practice. We propose a new FCN-based structure to extract the distinctive features of video regions. In general, entire frames are fed to the proposed FCN. As a result, features of all regions are removed in a systematic manner. The extracted and localized anomalies in the video is done so by carefully and methodically running through them. A standard NVIDIA TITAN GPU processes approx. 370 frames per second (fps) when analyzing (low-resolution) frames of size 320 × 240. This is considered to be "very fast". Convolution and pooling operations in CNNs are responsible for extracting regions from input data using a specific stride and size. In this paper, by analyzing the deep layers of output, we plan and put forward a unique procedure for localizing and detecting abnormal regions in a frame. The idea of localizing a receptive field is inspired by the faster-RCNN in [5]. Two Gaussian models are defined based on the description of all normal training regions. The first model is generated by the k th layer of the CNN, while the second model is based on its transformation by the (k + 1)th convolutional layer. If in the testing phase, there are regions which differ in the first Gaussian model, we have encoded them as being a confident anomaly. Those regions which fit completely to the first model are labeled as being normal. Because the remaining regions fall

 DOI : 10.35291/2454-9150.2020.0212

a little below the threshold with a very small difference; we evaluated them by the second Gaussian model and are indicated by a sparse autoencoder. The key points of this paper are as follows:

• To the best of our knowledge, this is the first time that an FCN is used for anomaly detection.

• We adapt a pre-trained classification CNN to an FCN for generating video regions to describe motion and shape concurrently.

• For efficiency in time anomaly localization and detection, a new FCN architecture is proposed.

• The proposed method performs as well as state-of-the-art methods, but our method outperforms those with respect to time; we have real-time for typical applications.

• We achieved a processing speed of 370 fps on a standard GPU; this is about three times faster than the fastest existing method reported so far

II. OBJECTIVE

The objective of this system is to distinguish between objects: persons, bags, vehicles etc. Video surveillance suspect detection system detects, identifies and tracks objects within existing CCTV systems and automatically detects suspicious behaviors and other violations of established security policies and procedures.

• An additional objective of this system is that it reduces people's cost with smart technology.

• It integrates systems with advanced surveillance technology.

• The main advantages of this system is it can accurately detect moving objects.

• The system can be implemented for real-time detection of objects.

• The confidentiality is maintained by restricting user and role based access.

• Sound alerts and alarms according to user escalation procedures.

III. LITERATURE SURVEY

1. **Motion Estimation for Human Activity Surveillance**
   This project is a video surveillance system that detects, tracks, and monitors suspicious activities. The advantage of this paper is that it can monitor multiple screens simultaneously without the disadvantage of dropping concentration. Elevates operational effectiveness and efficiency. However, the disadvantage is that it cannot differentiate between similar looking objects.
2. **Object detection and recognition by image parsing.**
   Illustrates importance of combining bottom up and top down models and performing segmentation and object detection simultaneously. The advantage is that the system is able to segment the images, detect faces. Disadvantage is that it cannot differentiate between a shadow and a real object.
3. **Automated real time detection of suspicious behavior in public transport areas.**
   This project is a video surveillance system based on object tracking. Based on real time detection of object. The drawback is that pattern matching
   algorithm is used which is based on depth first search.
4. **Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes.**
   This paper proposes a fast and reliable method for anomaly detection and localization in video data showing crowded scenes. This deep network operates on small cubic patches as being the first stage, before carefully resizing remaining candidates of interest, and evaluating those at the second stage using a more complex and deeper 3D convolutional neural network (CNN)
5. **Faster R-CNN: Towards real-time object detection with region proposal networks**

   In this work, we introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals An RPN is a fully-convolutional network that simultaneously predicts object bounds and objectless scores at each position.
6. **Simple deep learning baseline for image classification**
   In this paper, they propose a very simple deep learning network for image classification which comprises only the very basic data processing components: cascaded principal component analysis (PCA), binary hashing, and block-wise histograms. PCA is employed to learn multistage filter banks. It is followed by simple binary hashing and block histograms for indexing and pooling.
7. **You Only Look Once: Unified, Real-Time Object Detection.**
   This project is a video surveillance system based on object tracking. It is based on real time detection of objects. It struggles to precisely localize some objects, especially small ones.

IV. EXISTING SYSTEM

An unsupervised deep learning approach is used in [6] for extracting anomalies in crowded scenes. In this approach, shapes and features are extracted using a PCANet [7] from 3D gradients. Then, a deep Gaussian mixture model (GMM) is used to build a model that defines the event patterns. A PCANet is also used in [8]. In this study, authors exploit the human visual system (HVS) to define features in the spatial domain. On the other hand, a

 DOI : 10.35291/2454-9150.2020.0212

multiscale histogram of optical flow (MHOF) is used to represent motion features of the video. PCANet is adopted to exploit these spatio-temporal features in order to distinguish abnormal events. An object shows an anomaly if it does not follow learned normal trajectories. This approach has many disadvantage such as it cannot handle occlusions effectively, and is too complex for processing crowded scenes. To avoid these two weaknesses, it is proposed to use spatio-temporal low level features such as optical flow or gradients. **Information Loss:** Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information **Independent variables become less interpretable:** After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features CNN-based approaches outperformed state of-the-art methods in different areas including image classification, object detection or activity recognition's. .CNN proved recently to be useful for defining effective data analysis techniques for various applications. In spite of these benefits, CNNs are computationally slow, especially when considering block-wise methods. Major problems in anomaly detection using CNNs are as follows:

1. CNN works too slow for patch-based methods; thus CNN is considered as a time consuming procedure.
2. Since CNN training is fully a supervised learning method, the detection of anomalies in real-world videos suffers from a basic impossibility of training large sets of samples from non-existing classes of anomalies.

V. PROBLEM STATEMENT

Activity detection is a very crucial component of video surveillance systems for activity-based analysis. Traditionally, the video feed from CCTV cameras were analyzed by human operators. These operators monitor multiple screens at a time searching for anomalous activities. This is a very dull and inefficient way of monitoring. Inefficient because humans are prone to errors. A human operator cannot monitor multiple screens simultaneously. So, it becomes very difficult to obtain activity information quickly and accurately. This is precisely why we need an automated system and suspicious activity detection system has created the perfect solution. Video-based suspicious activity detection systems can replace or help human operators to monitor unusual activity. The system gives them an instant and accurate response.

VI. PROPOSED SYSTEM

In video data, all the abnormal events are defined in terms of irregular shapes or motion, or possibly a mix of both. As a result of this definition, identifying the shapes and motion is an essential task for anomaly detection and localization. So as to spot the motion properties of events, we would like a series of frames. In other words, a single frame does not include motion properties; it only provides shape information of that specific frame. For analyzing both shape and motion, we consider the pixel-wise average of frame It and previous frame It−1, denoted by I 0 t (not to be confused with a derivative),

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

where it is in the frame in the video. For detecting anomalies in It, we use the sequence Dt = hI 0 t−4 , I0 t−2 , I0 t i. .

We start with this sequence Dt when representing video frames on grids of decreasing size w×h. Dt is defined on a grid Ω0 of size w0×h0. The sequence Dt is subsequently passed on to an FCN, defined by the k th intermediate convolutional layer, for k = 0, 1, . . . , L, each defined on a grid Ωk of size wk ×hk, where wk > wk+1, and hk > hk+1. We use L = 3 for the number of convolutional layers. The output of the k th intermediate convolutional layer of the FCN are feature vectors fk ∈ R mk (i.e. each containing mk real feature values), satisfying mk ≤ mk+1, starting with m0 = 1. For the input sequence Dt, the output of the k th convolutional layer is a matrix of vector values:

$$\{f_k^t(i,j,1:m_k)\}_{(i,j)=(1,1)}^{(w_k,h_k)} = \left\{[f_k^t(i,j,1),\ldots,f_k^t(i,j,m_k)]^\top\right\}_{(i,j)=(1,1)}^{(w_k,h_k)} \quad (2)$$

Each feature vector f t k (i, j, 1 : mk) is derived from a specific receptive field (i.e. a sub-region of input Dt). In other words, first, a high-level description of Dt is provided for the t th frame of the video. Second, Dt is represented subsequently by the k th intermediate convolutional layer of the FCN, for k = 1, . . . , L. This representation is used for identifying a set of partially pairwise overlapping regions in Ωk, called the receptive fields. Hence, we represent frame It at first by sequence Dt on Ω0, and then by mk maps.

$$f_{k,l} = \{f_k^t(i,j,l)\}_{(i,j)=(1,1)}^{(w_k,h_k)}, \text{ for } l = 1,2,\ldots,m_k \quad (3)$$

on Ωk, for k = 1, . . . , L. Recall that the size wk × hk decreases with increases of k values. Suppose that we have q training frames from a video which are considered to be normal. To represent these normal frames with respect to the k th convolutional layer of the FCN (AlexNet without its fully connected layers), we have wk × hk × q vectors of length mk, defining our 2D normal region descriptions; they are generated automatically by a pre-trained FCN . For modeling the normal behavior, a Gaussian distribution is fitted as a one class classifier to the descriptions of normal regions so that it defines our normal reference model. In the testing phase, a test frame It is described in a similar way by a set of regional features. Those regions which differ from the normal reference model are labeled as being abnormal. In particular, the features generated by a pre-trained CNN (2nd layer of AlexNet) are sufficiently

 DOI : 10.35291/2454-9150.2020.0212

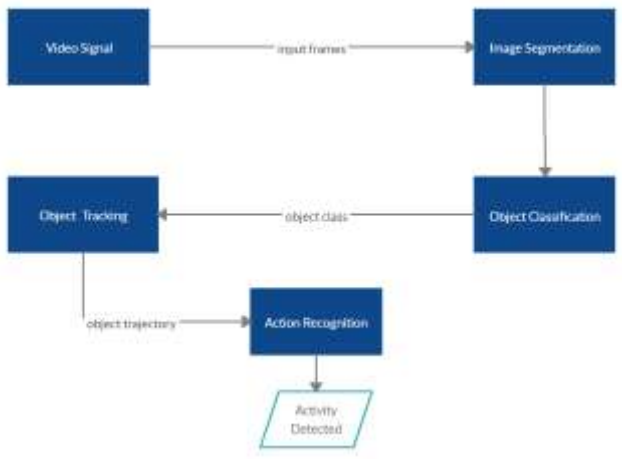
discriminative. These features are learned based on a set of independent images which are not necessarily related to video surveillance applications only.

In this paper, the video is represented employing a set of regional features. These features are extracted densely and their description is given by feature vectors in the output of the k th convolutional layer. See Equ. (2). Gaussian classifier G1(.) is fitted to all or any normal regional features generated by the FCN. Those regional features that their distance to G1(.) is greater than threshold α are considered to be abnormal. Those ones that are compatible to G1 (i.e. their distance is a smaller amount than threshold β) are labeled as being normal. A neighborhood is suspicious if it's a distance to G1 being between α and β. All suspicious regions are given to subsequent convolutional layer which is trained on all normal regions generated by the pre-trained FCN. The new representation of those suspicious regions is more discriminative and denoted by

$$T_{k,n} = \left\{T_k^t(i,j,n)\right\}_{(i,j)=(1,1)}^{(w_k^t, h_k^t)}, \quad \text{for } n = 1, 2, \ldots, h \qquad (4)$$

where h is the size of the feature vectors generated by the auto-encoder, which equals the size of the hidden layers. In this step, only the suspicious regions are processed. Thus, some points (i, j) in grid (wk, hk) are ignored and not analyzed in the grid (w 0 k , h0 k). Similar to G1, we create a Gaussian classifier G2 on all of the normal training regional features which are represented by our autoencoder. Those regions which are not sufficiently fitted to G2 are considered to be abnormal. Equations (5) and (6) summarize anomaly detection by using two fitted Gaussian classifiers.

VII. System Architecture



The camera will capture the video signal. The input of the video signal will be taken in the form of frames, which will then be segmented. The segmented images will be classified into objects, if any. The classified objects will be then tracked continuously, until they're in the video frame If there is any suspicious activity recognized by the system, in those objects that are being tracked, an alarm will be raised.

VIII. Conclusion

Thus, we have tried to implement "Activity Detection" using a pre-trained supervised FCN which is modified into an unsupervised FCN with the help of fully convolutional neural networks (FCNs) and temporal data, which in turn detects the anomalies in scenes. Many problems in computer vision were saturating on their accuracy before a decade.

However, with the increase of deep learning techniques, the accuracy of these problems drastically improved. The need this system arises due to various reasons which include: Due to increased crime rates around the world, many organizations are deploying video surveillance.

systems at their locations with CCTV cameras, which helps to prevent the threat before the crime actually happens because of the captured data.

References

[1] M. Sabokrou, M. Fayyaz, M. Fathy and R. Klette, "Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes," in IEEE Transactions on Image Processing, vol. 26, no. 4, pp. 1992-2004, April 2017.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Advances Neural Information Processing Systems (2012) 1097–1105.

[3] P. A. Dhulekar, S. T. Gandhe, A. Shewale, S. Sonawane and V. Yelmame, "Motion estimation for human activity surveillance," 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), Pune, 2017, pp. 82-85.

[4] M. Elhamod and M. D. Levine, "Automated Real-Time Detection of Potentially Suspicious Behavior in Public Transport Areas," in IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 2, pp. 688-699, June 2013

[5] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Analysis Machine Intelligence (2015), abs/1506.01497.

[6] F. Yachuang, Y. Yuan, L. Xiaoqiang, Learning deep event models for crowd anomaly detection, Neurocomputing (2017) 548–556.

[7] C. Tsung-Han, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, PcaNet: A simple deep learning baseline for image classification?, IEEE Trans. Image Processing (2015) 5017–5032.

[8] F. Zhijun, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, S. Chen, Abnormal event detection in crowded scenes based on deep learning, Multimedia Tools Applications (2016) 14617–14639.

 DOI : 10.35291/2454-9150.2020.0212