IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106076|

A Complete Convolutional Neural Network for Rapid Anomaly Detection in Crowded Scenes

G Nagi Reddy[#1], Shaik Irfan Babu[*2]

[1#]Department of Computer Science and Engineering, MGIT, Hyderabad, India

[2#]Department of Computer Science and Engineering, MGIT, Hyderabad, India

ABSTRACT: THE use of surveillance cameras requires that computer vision technologies need to be involved in the analysis of very large volumes of video data. The detection of anomalies in captured scenes is one of the applications in this area and it must deal with many challenges. The detection of abnormal behavior in crowded scenes must deal with many challenges Auto anomaly detection is very appreciable as it reduces human involvement and human work. Detecting anomalies is difficult as there is no proper definition for anomalies. Defining anomalies is not easy as their meaning changes with respect to the context they are present in, and it needs additional information to classify events as anomalies. Fully CNN model is used to detect anomalies in videos, by analyzing individual frames and regions in frames. In this proposed project, anomalies are defined as vehicles such as bicycle, van, car, and people doing skating, moving on wheelchair are considered. UCSD Dataset videos are used in anomaly detection. CAE and Gabor filter are used in implementing this project. This model represents feature representation and outlier detection. Experimental results suggest that detection of the proposed method outperforms existing methods in terms of accuracy. We can achieve a processing speed which is faster than the fastest existing method reported so far.

KEYWORDS: **CNN,** Histogram of Gradients, Histogram of Optic Flow, Gaussian mixture model, Markov random field

I. INTRODUCTION

The use of surveillance cameras requires that computer vision technologies need to be involved in the analysis of very large volumes of video data. The detection of anomalies in captured scenes is one of the applications in this area. In general, an event is considered to identify an “anomaly” when it occurs rarely, or unexpected. In the context of crowded scene videos, anomalies are formed by rare shapes or rare motions. For example, a video showing pedestrians, possible anomalies could be riding bicycle alongside pedestrians, any accident on pedestrians etc. Since looking for unknown shapes or motions is a time-consuming task; state-of-the-art approaches learn regions or patches of normal frames as reference models. Classifying these regions into normal and abnormal requires extensive sets of training samples to describe the properties of each region efficiently. There are numerous ways to describe region properties. Trajectory-based methods have been used to define behaviors of objects. Recently, for modeling spatio-temporal properties of video data, low-level features such as histogram of gradients (HoG) and histogram of optic flow (HoF) are used. These trajectory-based methods have two main disadvantages. They cannot handle occlusion problems, and they also suffer from high complexity, especially in crowded scenes.

This paper is implemented to detect anomalies in videos showing crowded scenes and is developed using python language. The main aim is to build a model that analyzes frames of videos and detects anomalies and used Fully CNN model to derive features from video frames. It takes video as input, analyses video frames region-wise and localizes anomalies, presents anomalies as output to user.

CNNs proved recently to be useful for defining effective data analysis techniques for various applications. CNN-based approaches outperform state of-the-art methods in different areas including image classification, object detection or activity recognition. Despite these benefits, CNNs are computationally slow, especially when considering block-wise methods. Thus, dividing a video into a set of patches and representing them by using CNNs, should be followed by a further analysis about possible ways of speed-ups.

The proposed system ‘Fully CNN for Fast Anomaly Detection in Crowded Scenes’ is implemented using Full CNN model. To overcome problems raised using CNN, we propose a new **FCN-based** structure to extract the distinctive features of video regions. In this approach, we are using Convolutional Autoencoder, which is very useful for semantic

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106076|

segmentation and noise reduction. We also use Gabor filter, which helps in edge detection. FCN's are unsupervised learning model. As we are using Autoencoder, which learns filters itself based on input and then extracts features. Being an unsupervised model, it is very efficient in anomaly detection. In general, entire frames are fed to the proposed **model.** Frames are analyzed regions wise and are compared with previous set of frames. As a result, features of all regions are extracted efficiently. Those regions which are found to vary from previous ones are consider anomaly. By analyzing the regions, anomalies in the video are extracted and localized. The processes of convolution and pooling, in all the CNN layers, run concurrently.

II. LITERATURE SURVEY

In circumstances of anomaly detection, object trajectory estimation is frequently of interest. If an object does not follow learned normal trajectories, it is said to be abnormal. This method has a number of flaws, including the inability to handle occlusions properly and being too sophisticated for processing busy situations. To overcome these two flaws, it is suggested that low-level spatiotemporal features such as optical flow or gradients be used. Zhang et al. (2005) employ a Markov random field (MRF) model to characterise the normal patterns of a movie in terms of rarity, unexpectedness, and importance. Boiman and Irani (2007) consider an occurrence as abnormal if its reconstruction is unachievable by using previous observations exclusively. For simulating the histograms of optical flow in local locations, Adam et al. (2008) adopt an exponential distribution.

Mahadevan et al. (2010) propose using a mixture of dynamic textures (MDT) to represent a video. The represented features are fitted into a Gaussian mixture model using this method. The MDT is expanded and explained in further depth by Li et al. (2014). For modelling local optical flow patterns, Kim and Grauman (2009) use a mixture of probabilistic PCA (MPPCA) models. They also employ an MRF to learn the standard patterns. Benezeth et al. present a method for behaviour modelling based on pixel motion attributes (2009). They described the video by learning a space-time co-occurrence matrix for regular events.

A Gaussian model is fitted to spatiotemporal gradient characteristics in Kratz and Nishino (2009), and a hidden Markov model (HMM) is utilised to detect aberrant occurrences. Mehran et al. (2009) propose social force (SF) as an effective technique for modelling aberrant crowd dynamics. Zaharescu and Wildes propose detecting abnormal behaviour using an approach based on spatial-temporal oriented energy filtering (2010). Cong et al. (2011) use normal data to create an over-complete normal basis set. If recreating a patch with this basis set is impossible, it is anomalous.

Antic et al. (2011) propose a scene parsing approach in Antic and Ommer (2011). Normal training explains all item hypotheses for the foreground of a frame. Hypotheses that cannot be explained by conventional training are called anomalous. Saligrama et al. propose a method based on clustering test data using optic-flow features in Saligrama and Chen (2012). Ullah and Conci (2012) proposed a crowd motion segmentation method based on a cut/max-flow algorithm. A flow is termed anomalous if it does not follow the regular motion model. A fast (140–150 fps) anomaly detection method based on sparse representation has been proposed by Lu et al. (2013).

In their work, Roshtkhari and Levine (2013a) adopt an adaptation of the bag of video words (BOV) approach. Zhu et al. (2013) describe a context-aware anomaly identification system in which the authors depict a video using motions and the context of the movie. A method for modelling both motion and shape with respect to a descriptor (dubbed "motion context") is proposed by Cong et al. (2013); the authors treat anomaly detection as a matching problem. Roshtkhari and Levine (2013b) present a method for constructing a hierarchical codebook to learn the dominant events of a movie. Ullah et al. (2013) use trained particles to train an MLP neural network to extract video behaviour. The behaviour of particles is learnt by utilising extracted information using a Gaussian mixture model (GMM). Ullah et al. (2014b) also present an MLP neural network for extracting corner characteristics from normal training samples, as well as labelling test samples with that MLP. Ullah et al. (2014a) use a random forest with corner features to extract corner features and analyse them based on their motion attributes using an enthalpy model. Based on hierarchical activity-pattern discovery, Xu et al. (2014) present a unified anomaly energy function for detecting anomalies.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106076|

III. METHODOLOGY

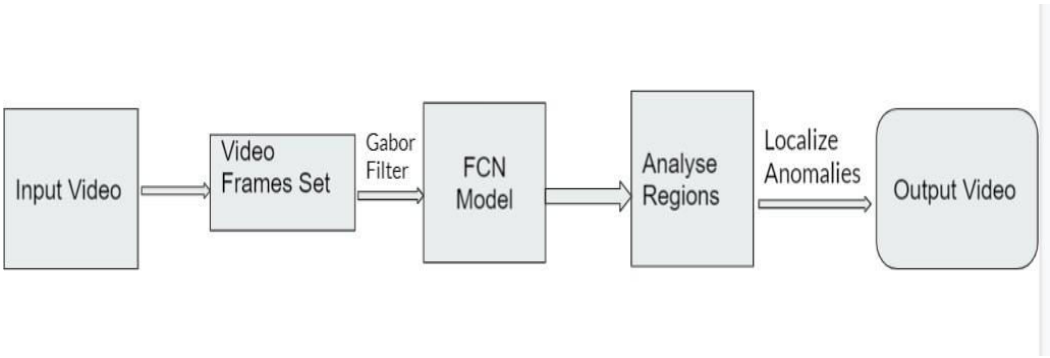

Figure 1: Schematic sketch of the proposed method

Figure 1. shows, Fully CNN (FCN) is a type of CNN, with modifications in model architecture which makes it an unsupervised model. FCN is a network that does not contain any "Dense" layers (as in traditional CNNs) instead it contains 1x1 convolutions that perform the task of fully connected layers (Dense layers). Though the absence of dense layers makes it possible to feed in variable inputs, there are a couple of techniques that enable us to use dense layers while cherishing variable input dimensions. FCN is learning filters everywhere. Even the decision-making layers at the end of the network are filters. FCN tries to learn representations and make decisions based on local spatial input. Appending a fully connected layer enables the network to learn something using global information where the spatial arrangement of the input falls away and need not apply. FCNs are an architecture used mainly for semantic segmentation. They employ solely locally connected layers, such as convolution, pooling, and up sampling. FCNs can be used for image classification, images semantic segmentation and high/low-level features extraction from images. FCN can also be applied to frames of videos to extract features from video frames. In this paper, used of Convolutional Autoencoder, a type of Fully CNN model, which contains only convolutional layers and is a unsupervised model. It can effectively extract features from frames of videos and is proved to be quite successful for anomaly detection. Along with CAE, we also use Gaussian Classifier, which makes use of gaussian probability distributions for making classifications. This classifier helps in deciding whether a region of video frame is anomaly or normal. The output of Gaussian classifier is analysed along with the Euclidian distance, which shows how far this region is from normal regions.

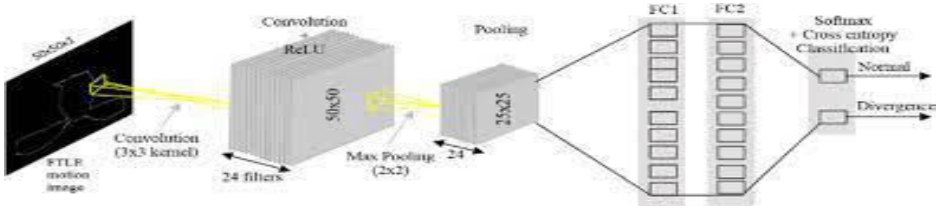

Figure 2: Architecture of Convolutional Neural Network

The first convolutional layer has m1 kernels of size x1 × y1. They are convolved on sequence Dt for considering the tth frame. As a result of this convolution, a feature is extracted. Recall that each region for the input of the FCN is described by a feature vector of length m1. In this continuous process, we have mk maps as output for the kth layer. Consequently, a point in the output of the kth layer is a description for a subset of overlapping (x1 × y1)th receptive fields in the input of the FCN. See Fig. 1. The order of layers in the modified version of AlexNet is denoted by

AlexNet Order → [C_1, S_1, C_2, S_2, C_3, fc_1, fc_2] (1)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106076|

where C and S are a convolutional layer and a sub-sampling layer, respectively. The two final layers are fully connected. Assume that n regional feature vectors (i1, j1) ⋯(in, jn), generated in layer Ck on grid Ωk, are identified as showing an anomaly. The location (i, j) in Ωk corresponds as the rectangular region in the original frame.

Suppose we have m_k kernels of size xk × yk which are then convolved with stride d on the output of the previous layer of Ck. − C ij (,) k 1 is the (rectangular) set of all locations in Ωk−1 which are mapped in the FCN on (i, j) in Ωk. Function −Sk 1 is defined in an analogous way. The sub-sampling (mean pooling) layer can also be considered as a convolutional layer which has only one kernel. Any region, detected as being an abnormal region in the original frame (i.e., in Ω0), is then a combination of some overlapping and large patches. This leads to a poor localization performance. As a case in point, a detection in the 2nd layer causes 51 × 51 overlapping receptive fields. To achieve more accuracy in anomaly detection, those pixels in Ω0 are identified to show an anomaly which are covered by more than ζ related receptive fields (we decided for ζ=3 experimentally).

IV. RESULTS

Figure 3: Output of the proposed method on Car, Wheelchair and Skating

To evaluate and compare the experimental results, used different datasets. In Figure 3, Dominant dynamic objects in this dataset are Car, Wheelchair and Skating where crowd density varies from low to high. An appearing object such as a car, skateboarder, wheelchair, or bicycle is considered to create an anomaly. All training frames in this dataset are normal and contain pedestrians only. This dataset has 12 sequences for testing, and 16 video sequences for training, with 320 × 240 resolution. For evaluating the localization, the ground truth of all test frames is available. The total numbers of abnormal and normal frames are ≈ 2384 and ≈ 2566, respectively. Subway (Adam et al., 2008). This dataset contains two sequences recorded at the entrance (1 h and 36 min, 144,249 frames) and exit (43 min, 64,900 frames) of a subway station. People entering and eating the station usually behave normally. Abnormal events are defined by people moving in the wrong direction (i.e., exiting the entrance or entering the exit), or avoiding payment. This dataset has two limitations: The number of anomalies is low, and there are predictable spatial localizations (at entrance or exit regions).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106076|

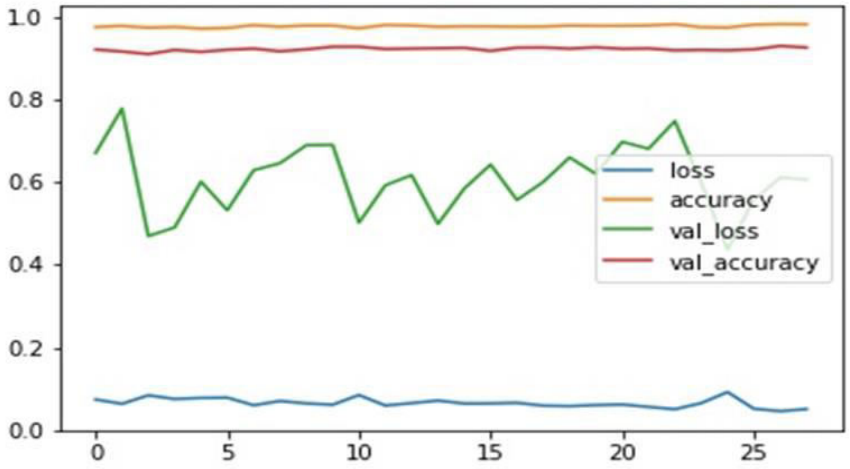

Figure 4: Validation Accuracy

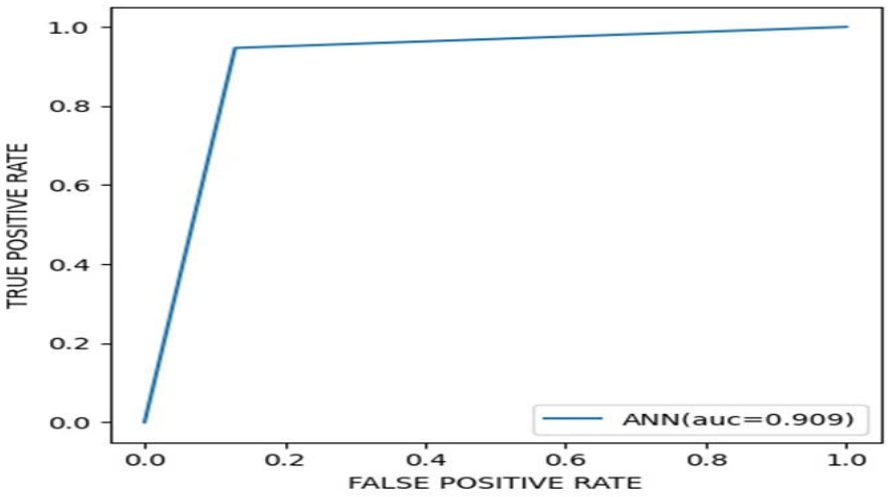

Figure 5: Roc Curve

Figure 4. shows that model error of the validation testing. Based upon this validation testing we should know the accuracy of our model. In above graph training accuracy is 98%, validation accuracy is 91%. To reduce the validation error used different types of regulation techniques are used. Figure 5. shows maximum AUC (92.9%) measures the model performance.

V. CONCULSION

A new FCN architecture for creating and describing aberrant regions in videos is presented in this paper. The regional characteristics are context-free by utilising the strength of the FCN architecture for patch-wise actions on input data. Furthermore, the suggested FCN is a hybrid of a pre-trained CNN and a new convolutional layer in which kernels are trained using the training video. The suggested FCN's final convolutional layer must be taught. In terms of processing speed, the proposed method outperforms existing methods. It's also a way to get around the restrictions of the training samples needed to learn a complete CNN. This strategy allows us to perform a deep learning-based method at around 370 frames per second. Overall, the proposed method for finding strange things in video data works quickly and correctly.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106076|

REFERENCES

[1] Mohammad Sabokrou , Mohsen Fayyaz , Mahmood Fathy , Zahra. Moayed, Reinhard Klette , "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes", ELSEVIER, Computer Vision and Image Understanding ,Volume 172, July 2018, Pages 88-97

[2] Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D., 2008. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. 30 (3), 555–560.

[3] Cong, Y., Yuan, J., Liu, J., 2011. Sparse reconstruction cost for abnormal event detection. Computer Vision Pattern Recognition. pp. 3449–3456.

[4] Antic, B., Ommer, B., 2011. Video parsing for abnormality detection. International Conference on Computer Vision. pp. 2415–2422.

[5] Antonakaki, P., Kosmopoulos, D., Perantonis, S.J., 2009. Detecting abnormal human behaviour using multiple cameras. Signal Process. 89 (9), 1723–1738.

[6] Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D., 2008. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. 30 (3), 555–560.

[7] Giusti, A., Ciresan, D.C., Masci, J., Gambardella, L.M., Schmidhuber, J., 2013. Fast image scanning with deep max-pooling convolutional neural networks. IEEE International Conference on Image Processing. pp. 4034–4038.

[8] Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Computer Vision Pattern Recognition. pp. 580–587.

[9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. Computer Vision Pattern Recognition. pp. 248–255.

[10] Cong, Y., Yuan, J., Tang, Y., 2013. Video anomaly search in crowded scenes via spatiotemporal motion context. IEEE Trans. Inf. Forensics Secur. 8 (10), 1590–1599. MIT places database, 2017.

[11] Benezeth, Y., Jodoin, P.M., Saligrama, V., Rosenberger, C., 2009. Abnormal events de tection based on spatio-temporal co-occurrences. Computer Vision Pattern Recognition. pp. 2458–2465.

[12] Bertini, M., Bimbo, A.D., Seidenari, L., 2012. Multi-scale and real-time non-parametric approach for anomaly detection and localization. Comput. Vis. Image Underst. 320–329.

[13] Calderara, S., Heinemann, U., Prati, A., Cucchiara, R., Tishby, N., 2011. Detecting anomalies in people's trajectories using spectral graph analysis. Comput. Vis. Image Underst. 115 (8), 1099–1111.

[14] Cheng, K.W., Chen, Y.T., Fang, W.H., 2015. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. Computer Vision Pattern Recognition. pp. 2909–2917.

[15] Boiman, O., Irani, M., 2007. Detecting irregularities in images and in video. Int. J. Comput. Vis. 74 (1), 17–31.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details