Intelligent Security and Surveillance System Using Deep Learning Techniques

[1]Sujit N., [2]Dylan D., [3]Akshaye N., [4]Dr. Jagruti S.

[1,2,3]Student, [4]Professor, Fr. CRCE, Bandra(W), Mumbai ,India,

[1]sujitnoronha@outlook.com, [2]dsouzadylan2000@gmail.com, [3]akshaye.nair@gmail.com,

[4] jsave@frcrce.ac.in

Abstract: Surveillance is the monitoring and management of people's behavior in the corresponding surrounding. This may include remote surveillance using automated equipment such as closed-circuit television (CCTV) or electronic data interception such as Internet traffic. Governments use surveillance for intelligence collection, crime prevention, the security of a method, individual, entity, or object, and crime investigation. Deep learning based neural networks interpret video outputs from surveillance cameras to identify pedestrians, cars, objects, and events are used in artificial intelligence for video surveillance. AI is a rapidly growing phenomenon that has piqued the interest of experts from a wide range of disciplines as a means of facilitating complex problem-solving in ways that were previously impossible. This novel solution aims to create artificial intelligence that reduces the workload on security personnel by detecting different types of crimes happening in real-time on a particular scene and informing authorities of the same. It can also detect the number of people present in the scene and detect the faces of people who have previous criminal records.

Keywords — Artificial Intelligence, Activity Recognition, Convolutional Neural Network, Deep Learning, Social Distancing, Surveillance Videos

I. INTRODUCTION

The use of surveillance cameras in public places leads to ensuring public safety. Not only do surveillance cameras help in Post Crime investigations but they also act as deterrents. CCTV systems currently deployed over the globe need a lot of human intervention and manpower to monitor the video outputs. The personnel monitoring the CCTV systems will be inefficient and ineffective when monitoring/dealing with a lot of video outputs over a prolonged period of time causing fatigue and stress in people monitoring the systems. Homicide accounts for less than 1% of global deaths, but it accounts for up to 10% in some nations. Homicide claimed the lives of 0.7 percent of people worldwide in 2017. Homicide was responsible for less than 0.1 percent of deaths in most of Western Europe, for example. It was less than 0.5 percent in most of Eastern Europe, North Africa, Asia, and Oceania. It was 0.7 percent in the United States. However, in some nations, the situation is different. An intelligent surveillance system does not act as a physical barrier or limit access to steal, or a person more difficult to assault and rob. Although CCTV-based systems have many functions the primary function is to trigger a perceptual mechanism in a potential criminal. And establishes a perception that if he commits a crime, he will be caught. The other benefits include information gathering, aid to police investigations, and a sense of safety among the people.

II. LITERATURE SURVEY

Detecting inconsistencies in surveillance images is one of the most critical tasks in artificial intelligence and computer vision, although there have been attempts made to solve this difficult task[3]. The paper Real-world Anomaly Detection in Surveillance[1] Videos proposes a new system that is used to detect anomalous activity in the surroundings such as violence, arson, fire, shooting, robbery, assault, explosion, etc. thus providing an enhanced and effective way to classify videos. Yang and Dongfang[10] have given a comprehensive study of how monitoring social distancing and sending alerts to authorities in case of violations could help controlling overcrowding in places. Image classification is a critical task in deep learning that is used in a wide range of applications and has a high usability and scope. MobileNet is a CNN-based architecture for Image Classification. The paper[6] titled "MobileNetV2: Inverted Residuals and Linear Bottlenecks" has described a modern mobile architecture that enhances mobile models' state-of-the-art efficiency on multiple tasks and benchmarks, as well as across a range of model sizes. The MobileNetV2

 DOI : 10.35291/2454-9150.2021.0168

architecture is based on an inverted residual structure, in which the residual block's input and output are thin bottleneck layers, and MobileNetV2 filters features in the intermediate expansion layer with lightweight depth-wise convolutions, in contrast to traditional residual models, which use extended representations in the input.. Tremendous strides have been made in face detection. The paper titled "FaceNet: A Unified Embedding for Face Recognition and Clustering"[4] The network used in FaceNet is capable of mapping faces into a complex Euclidean space, using distance to quantify similarities . This approach is based on learning a Euclidean embedding of each face by the use of a deep Convolutional Neural Network, the neural network is then equipped to specifically correlate with facial similarities in the embedded space with the L2 distances. The loss function used in this system is known as triplet loss and is used in the process of training this architecture. Batch normalization is a deep neural network technique that reduces the impact of unstable gradients. The YOLO framework is a state-of-the-art object detection model. The paper[5] gives us more information on how YOLO trains on frames of videos and enhances detection speed and efficiency. Compared to conventional object detection approaches, this unified model has many advantages. On testing the YOLO v4 on the MS COCO dataset, the performance of the system in real time were as follows, on a Tesla V100 it achieved 65 FPS with a 43.5 percent AP.

III. PROPOSED SYSTEM

To improve the detection capabilities of the existing solution we propose to use state-of-the-art algorithms in the field of computer vision and deep learning for detecting and monitoring people/objects in the surroundings. Our proposed system works in the following way as shown in Fig 1:

1.) The CCTV will capture every footage and keep a record of it if required in the future. This footage is passed through our system where the video is broken down into frames and each frame is examined thoroughly.

2.) These broken frames are passed through the neural network(CNN) where first it checks for social distancing, if not it gives an alert that people are not following the norms. Along with that, the frames are classified according to different events.

3.) Our system consists of five to six event classes viz. Assault and Fighting, Shooting, Road Accident, Explosion, Normal Event. For e.g. If there is an accident that occurred, then it will notify the hospital or if a property or a building is on fire, it will notify the fire station.

Use of object detection: Object recognition is a computer vision technique for detecting objects in photographs or videos. To generate meaningful results, object detection algorithms use machine learning or deep learning. Object detection will be used to detect people as well as any other objects that were included in the training dataset.

Activity recognition: An activity recognition system will also be implemented to determine what is happening in the video stream as a whole. If fights/ brawls take place the police can be notified, and thus necessary action can be taken. Activity recognition will be implemented using recurrent neural networks to provide analysis of the video over time.

Social Distance Detector: To help with ensuring social distancing is maintained in the neighborhood, we have created a Deep Learning based social distancing detector using object detection and classification that can detect if people are keeping a safe distance from each other by analyzing real-time video streams from the camera.

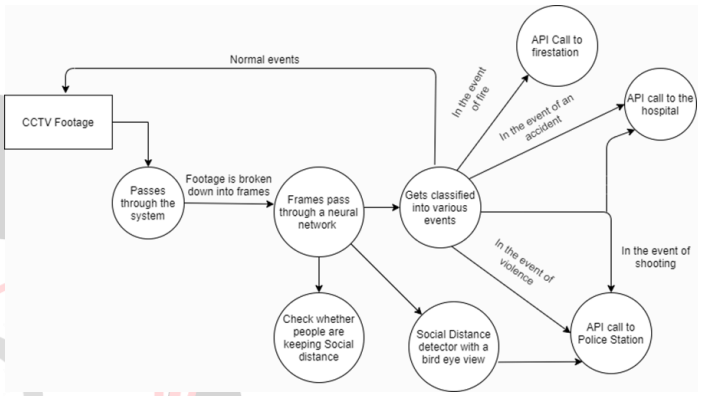

Fig 1. System Flow Diagram

IV. DATASET

We have created a dataset which consists of five classes namely:

Assault and Fighting: This event contains videos that show a person getting attacked by someone. If the CCTV captures people fighting, our system will immediately contact the police station through the API.

Shooting: This event contains videos showing the act of shooting someone with a gun. If the CCTV captures a person or people using guns, our system will send an alarm or notification to the police station.

Road Accident: This event contains videos of car or bike accidents. If these events occur and are captured by the CCTV, our system will immediately notify the hospital as well as the police station.

Explosion: This event contains videos of the bomb blast, short-circuit, etc. If the CCTV captures (for eg.) a building or a property on fire, it will immediately notify the nearest fire station.

Normal Event: This subset of images includes pedestrians walking, talking to each other, and engaging in social activities. Thus helping the model to differentiate between

 DOI : 10.35291/2454-9150.2021.0168

other dangerous or harmful behavior

The dataset has been created manually and contains almost 4000 pictures i.e. approximately 600 per class.

Fig 2. Sample Images of the Dataset

V. IMPLEMENTATION OF METHODOLOGY

A. *Creating our own dataset*

Training of videos from crime detection on a CPU would have been a herculean task. We prepare our own dataset by splitting up the videos from the UCF-Anomaly Detection Dataset into particular frames which help us train our model and give us faster inference. We have also taken videos from various other sources to help us with our inference.

B. *Activity Recognition*

Detecting suspicious activity using traditional methods is difficult and requires a lot of manpower compared to an automated system. The human operator could develop fatigue and lose concentration over a prolonged period of time reducing the effectiveness. In our system the real world input video is first split up into frames and then the activity is classified as whether it is a normal activity or a suspicious activity. We have used the MobileNet v2 based classification architecture to classify the frames into particular activities or crimes[6]. The MobileNet architecture gave us the best result among the many architectures that we had used. On completing inference on the particular frame the prediction is pushed into an array of 60 elements where an average is taken of all the elements in the array to predict the most common prediction at the particular point of time.

C. *Face Recognition*

For the detection and recognition of faces present on the current criminal database, we used the Facenet based architecture[4]. The network used in FaceNet is capable of mapping faces into a complex Euclidean space, using distance to quantify similarities . This approach is based on learning a Euclidean embedding of each face by the use of a deep Convolutional Neural Network, the neural network is then equipped to specifically correlate with facial similarities in the embedded space with the L2 distances. Thus, a lesser distance correlates to the same face while a large distance correlates to a different face.

D. *Social Distance Detector*

Object detection is a method for locating semantic objects in video scenes such as humans, animals, and carriages. The primary goal of a video surveillance system is to solve a variety of issues such as object detection and tracking. We have used the YOLO V-4 tiny to detect people in a particular scene[5]. We can also retrieve the number of people present in the scene using this technique and it will be shown to the client. YOLO stands for You Only Look Once, and it is a real-time object detection system that recognizes multiple objects in a single enclosure. It also recognizes objects faster and more precisely than other recognition systems.. Using Object Detection we can check whether Social Distancing is being maintained by people present in the scene or not. As not maintaining social distancing is also another type of crime we have the Euclidean Distance formula to calculate the distance between two people. The length of a line segment between two points in Euclidean space is known as the Euclidean distance in mathematics. If Social Distancing is not maintained for a particular period of time then authorities could be alerted, and strict action can be taken

E. *Usage Process Flow*

The final operation flow is shown in Fig 2. The system begins with the user/client feeding in the CCTV input to our model. The model detects the type of crime being committed in the scene in real-time and the respective authority is informed about it. If no crime is committed in that particular scene it then it classifies as a Normal Scene. Using object detection bounding boxes are drawn on people present in the scene, if the social distance is not maintained then the green box is turned to red. These results are sent to our client-side which gives us the number of people present in the scene, location of the scene, and the status of the scene. A graph shows the number of people present in a particular scene. If people present in the scene have past criminal records and are stored in the database then their faces will be detected automatically using Face Recognition and detection

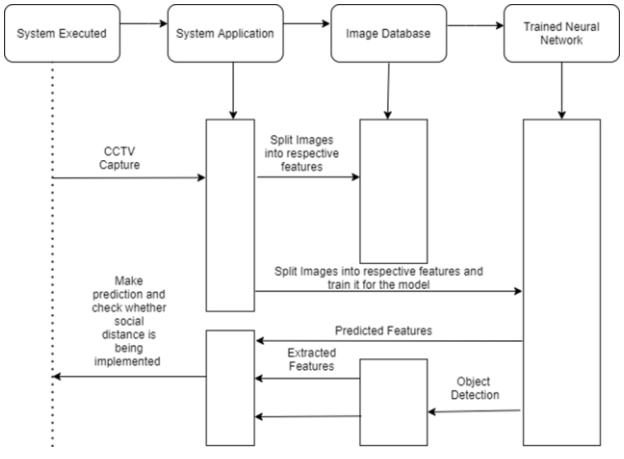

Fig 3. Usage Flow Diagram

 DOI : 10.35291/2454-9150.2021.0168

VI. Result And Analysis

The dataset was created using 1200 images, containing images from various sources. The training and validation images were split into an 80:20 ratio. The Mobilenet based model used a method known as transfer learning in which existing ImageNet weights were retained and additional layers were added to allow the model to fit the created dataset. The images are then resized and to a 255*255*3 sized vector and data augmentation is used to increase the diversity of the data available. Fig 4. provides the details related to the training and validation loss. The training and validation loss for the training and validation dataset is 0.1377 and 0.6985 respectively and the accuracy of the Mobile Net model is 96%. The YOLO-based object detector is pre-trained and uses the COCO datasets weights thus achieving state-of-the-art results on detecting people. The video inputs are fed into the models allowing the system to detect scene information, number of people, and social distancing among the people.

Fig. 5 demonstrates the system predicting a normal event. The probability box towards the left of the image indicates the probabilities of the event in the particular scene. Fig. 6 shows two people fighting in the video and the network predicts violence as the event in the scene. The prediction of the frame is then transferred to a REST-based backend application hosted on a private server which then saves the data to a database that allows the data to be displayed using a front-end for the authorities. Thus improving the efficiency and accuracy compared to traditional human operators.

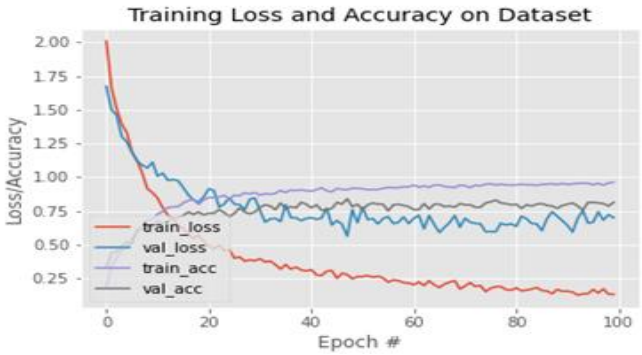

Fig 4. Training Loss

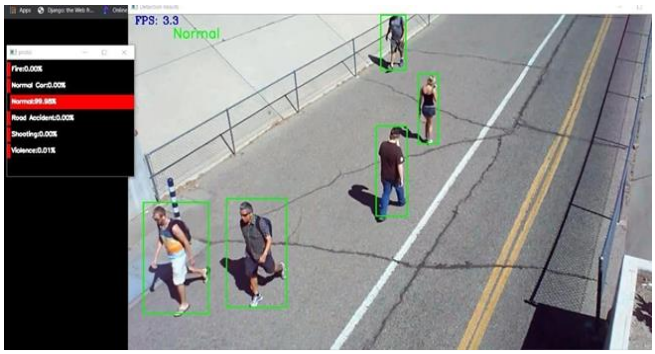

Fig 5. Prediction of a Normal Event

Fig 6. Prediction of a Violent Event

VII. Conclusion

This paper demonstrates the capabilities of a CNN based neural network for surveillance based applications. In addition, it provides better performance and requires lower training time compared to the existing LSTM-CNN based solutions[1]. Thus, improved performance and light-weight models help in running the application in real time on devices with low compute power rather than only on GPUs. Owing to better availability and wider use by reducing the costs of the surveillance systems computing hardware. The facial recognition system[4] provides accurate results in recognizing people in the particular frame allowing the respective authorities to scan through all people in crowded regions effectively.

VIII. Future Work

There is a lot of scope for future development as mentioned in the conclusion. The inclusion of a facial recognition system by only using the eyes will help detect criminals in a better way. The system's accuracy can also be improved by adding lighter object detection models to improve accuracy.

IX. References

[1] W. Sultani, C. Chen and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 6479-6488, doi: 10.1109/CVPR.2018.00678

[2] Ioffe, Sergey & Szegedy, Christian. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

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

[4] Schroff, Florian & Kalenichenko, Dmitry & Philbin, James. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. Proc. CVPR.

[5] Bochkovskiy, Alexey, Chien-Yao Wang and H. Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection." *ArXiv* abs/2004.10934 (2020):

 DOI : 10.35291/2454-9150.2021.0168

[6] MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Andrew G. Howard Menglong Zhu Bo Chen Dmitry Kalenichenko

[7] Lecun, Yann & Bottou, Leon & Bengio, Y. & Haffner, Patrick. (1998.Gradient- Based Learning Applied to Document Recognition. Proceedings of the IEEE. 86. 2278 - 2324. 10.1109/5.726791.

[8] Mishra, Pawan & Saroha, Gyanendra. (2016). A Study on Video Surveillance System for Object Detection and Tracking.

[9] https://docs.openvinotoolkit.org/latest/index.html

[10] Yang, Dongfang & Yurtsever, Ekim & Renganathan, Vishnu & Redmill, Keith & Ozguner, Umit. (2020). A Vision-based Social Distancing and Critical Density Detection System for COVID-19.

[11] Vishwakarma, Sarvesh & Agrawal, Anupam. (2012). A Survey on Activity Recognition and Behavior Understanding in Video Surveillance. The Visual Computer. 29. 10.1007/s00371-012-0752

[12] Owayjan, Michel & Dergham, Amer & Haber, Gerges & Fakih, Nidal & Hamoush, Ahmad & Abdo, Elie. (2013). Face Recognition Security System

[13] Zhao, Zhong-Qiu & Zheng, Peng & Xu, Shou-Tao & Wu, Xindong. (2019). Object Detection With Deep Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems. PP. 1-21. 10.1109/TNNLS.2018.2876865.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Computation 9(8) (November 1997),1735–1780.

[15] Srivastava, Nitish & Hinton, Geoffrey & Krizhevsky, Alex & Sutskever, Ilya & Salakhutdinov, Ruslan. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 15. 1929-1958.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778

[17] Cristani, Marco & Del Bue, Alessio & Murino, Vittorio & Setti, Francesco & Vinciarelli, Alessandro. (2020). The Visual Social Distancing Problem. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3008370.

[18] Xu, Dan & Ricci, Elisa & Yan, Yan & Song, Jingkuan & Sebe, Nicu. (2015). Learning Deep Representations of Appearance and Motion for Anomalous Event Detection. Computer Vision and Image Understanding. 10.1016/j.cviu.2016.10.010.

[19] LI, Wei-Xin & Mahadevan, Vijay & Vasconcelos, Nuno. (2013). Anomaly Detection and Localization in Crowded Scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 36. 18-32. 10.1109/TPAMI.2013.111.

[20] Mohammadi, Sadegh & Perina, Alessandro & Kiani, Hamed & Murino, Vittorio. (2016). Angry Crowds: Detecting Violent Events in Videos. 9911. 3-18. 10.1007/978-3-319-46478-7_1.

 DOI : 10.35291/2454-9150.2021.0168