INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 04 | APRIL - 2023 IMPACT FACTOR: 8.176 ISSN: 2582-3930

NAYAN NIRIKSHAN -

Violence Detection Using Surveillance Camera

Subhas B S
CSE dept.
Rajiv Gandhi Institution of Technology
Bangalore, India
subhashbs36@gmail.com

Sumanth Hegde
CSE dept.
Rajiv Gandhi Institution of Technology
Bangalore, India
sumanthdh13@gmail.com

Nityanand Kumar
CSE dept.
Rajiv Gandhi Institution of Technology
Bangalore, India
nityanandchoudhary420@gmail.com

Harsha R
CSE dept.
Rajiv Gandhi Institution of Technology
Bangalore, India
harsharharry@gmail.com

Bhagyashri Wakde
Assistant Professor Dept. of CSE
Rajiv Gandhi Institution of Technology
Bangalore, India
bhagyashelke2015@gmail.com

I. ABSTRACT

MOVE-NET POSE ESTIMATION IS A METHOD FOR REAL-TIME POSTURE AND MOVEMENT DETECTION OF THE HUMAN BODY USING DEEP LEARNING ALGORITHMS. MOVE-NET POSE ESTIMATION CAN BE USED TO EXAMINE THE VIDEO STREAM AND FIND ANY UNUSUAL OR VIOLENT BEHAVIOUR WHEN DETECTING VIOLENCE USING SECURITY CAMERAS. A SERIES OF KEY POINTS INDICATING THE POSITION OF JOINTS AND LIMBS ARE PRODUCED BY THE SYSTEM, AND THESE KEY POINTS CAN BE USED TO RECOGNIZE PARTICULAR MOVEMENTS AND POSTURES THAT ARE SUGGESTIVE OF VIOLENCE. THIS ENHANCES THE OVERALL SECURITY OF THE MONITORED LOCATION BY ENABLING THE REAL-TIME DETECTION AND RESPONSE TO VIOLENT SITUATIONS. ADDITIONALLY, MOVE-NET POSE ESTIMATION CAN BE TAILORED TO RECOGNIZE PARTICULAR POSTURES AND MOTIONS IN VARIOUS SETTINGS AND CONNECTED WITH OTHER SECURITY SYSTEMS LIKE ALARMS AND ACCESS CONTROL. BUT THERE ARE CONSTRAINTS.

***Index Terms*—**
Move-Net Lightning, Yolo v3, Single-Pose Detection, Convolutional Neural Network, Violence and Non-violence activity (CNN).

INTRODUCTION

In both public and private places, video monitoring has become a standard technique for maintaining security. Real-time activity monitoring and analysis has been simpler and more affordable because to developments in digital video recording, network video transmission, and video analytics. The huge amount of video footage produced by these systems, however, may be manually monitored, although this can be time-consuming, resource-intensive, and prone to human mistake. As a result, algorithms for computer vision and deep learning have been created, enhancing the effectiveness and precision of video surveillance systems by automating the monitoring process.

Movenet Pose Estimation, a deep learning-based tool that tracks human posture and movement in real time, is one such example. As a tool for violence detection employing security cameras, this technology offers a quick and efficient approach to recognise and respond to violent situations. The applications of Move-net Pose Estimation for violence detection are covered in detail in this study, along with a description of the system's operation, advantages and disadvantages, and possible integrations. In addition, privacy issues associated with video monitoring will be reviewed, and the affordability of Move-net Pose Estimation as a violence detection system will be examined.

Move-net Pose Estimation examines video feeds from security cameras in the context of violence detection to spot any unusual or aggressive activity. In order to identify specific motions and postures linked to violence, the system generates a collection of key points that reflect the position of joints and limbs. This makes it possible to identify and respond to violent situations in real time, improving the monitored area's overall security. Move-net Pose Estimation is a security system that can be combined with other security systems like alarms and access control and can be configured to identify movements and postures relevant to various situations.

Large datasets of human body movements and postures were used to train the deep learning algorithms employed by Move-net Pose Estimation. Move-net Stance Estimation is a trustworthy technique for violence detection since these algorithms are resistant to changes in lighting, camera angle, and body pose. By eliminating the need for human security employees and offering real-time alerts and responses to violent occurrences, the adoption of Move-net Pose Estimation can also offer a cost-effective alternative to conventional security methods.

However, privacy issues are brought up by the deployment of video surveillance systems, such as Move-net Pose Estimation. Individuals' privacy and the possibility of misuse are called into issue by the gathering and storing of video footage of them. The

ethical and legal ramifications of video monitoring must be taken into account in order to guarantee privacy rights.

In summary, Move-net Pose Estimation is an effective method for detecting aggression in video surveillance systems. It can increase the security of monitored locations and offer a cost-effective substitute for conventional security measures due to its capacity to automate the monitoring process and detect violent situations in real-time. Although privacy issues must be taken into account, using Move-net Pose Estimation can be a useful strategy for fostering safety and security.

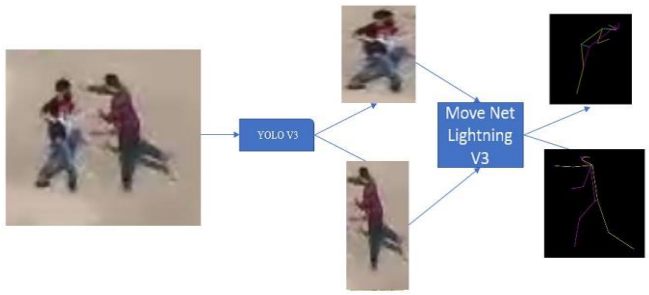
II. RELATED WORK

According to [01] a method of following an object's motion

Inside a scene. [03]The algorithms provide real-time identifications that are accurate, exact, and ideal for real-time traffic applications.[04] The systems and procedures for recognising, finding, and managing the stock of sterilised medical devices that are kept inside a sterilisation facility are the subject of this invention.[02] Next, we give a quick rundown of the different applications where real-time video processing techniques may be utilised to manage, watch over, or regulate traffic.[05] In this study, we build and assess feature sets to determine the significance of various body parts and traits for recognition.[06] The 21st Century: Behaviour, Performance, and Effectiveness. According to [07] we investigate the issue of feature sets for reliable visual object recognition using a test case of linear SVM-based person detection.

According to [09], there are two components to the visual tracking method: a motion model and an observation model. A motion model predicts the state in which the item will be based on the description of the preceding state. For example, these filters include the kal man [10] and particle [11] and [12] filters. The observation model, which confirms predictions for each frame, is a representation of the appearance information for the monitored object [13]. The observation model, as opposed to the motion model, is more crucial for visual tracking, claims the [08] research. Nowadays, generative and discriminative techniques are the two primary categories of observation models being used 11. Whereas generative techniques are more focused on identifying ROI that are similar to the desired item through template matching, dis-criminative approaches are more concerned with classification and working to separate the object from background [10][14][15][16]. Depending on how the object tracker is meant to be used, In contrast to generative CNN trackers, discriminative CNN trackers are often quicker and more accurate [09].

Deep visual tracking employs more than just CNNs. As they can establish temporal links between recognised items from Yolo, they are then clipped and sent to Move-net for posture prediction of the violent behaviours. Based on their pose estimation, Move-net will determine whether or not the activities are violent and preserve memories of earlier ones, recurrent neural networks are highly suited for sequence modelling. Given the success in the domains of handwriting [09] [11] and speech recognition [12], several authors propose combining RNNs with correlation filters [14] or even incorporating features created by RNN into CNN to



build a robust feature representation [14].Deep neural networks can be used for the selection of the best candidate towards tracking object in addition to extracting their attributes. In this case, there are two sorts of networks: feature extraction networks (FENs) and end-to-end networks (EENs). FENs have the advantage of obtaining high-quality features, which are then employed by conventional methods to learn model appearance, as opposed to EENs, which perform both tasks. As a result, EENs can offer a bounding box, segmentation representation, or, as a comprehensive solution, a probability map as a means of locating an object [15] [16].

Due to the nature of the sport, writers have chosen to concentrate on solutions that will offer quick solutions with the highest level of precision in situations when quick and fast detection is required with constant multiple object movement in different directions. To handle object detection and simple online and real-time tracking, Yolo [17] was used.

III. DESCRIPTION

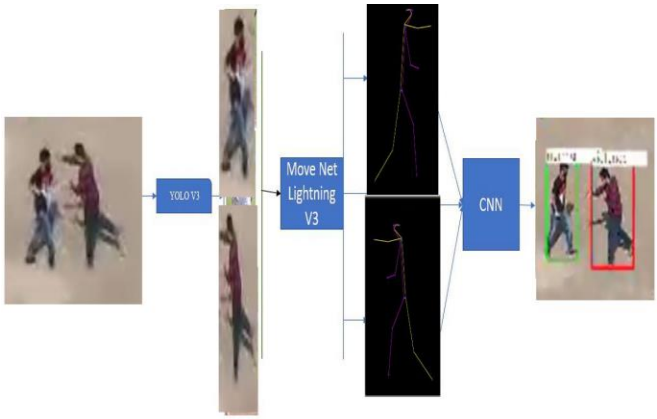
A. Yolo V3 and Move net lightning

Yolo V3 is a real-time data classification system for computer vision that identifies objections. Yolo version 3 is created by implementing 1 x 1 detection with kernels into feature maps of three different sizes at three distinct locations throughout the network.

Yolo V3 has been utilised in this work to recognise people, and the video is transferred from frame to frame. Yolo analyses each frame and outputs the bounding box for each character that appears in it. The photos are cropped using the bounding box coordinates before being sent for image pre-processing, where a black image with the same image dimensions is produced.

A position identification method called Move-net Lightning is used to collect 17 crucial human anatomy points. Here, we avoid having the algorithm take other people in the frame into account by sending each person's cropped image separately. The previously created black graphic contains the 17 important spots that have been identified and plotted. The 17 key points are as

follows: Nose 0; Left Eye 1; Right Eye 2; Left Ear 3; Right Ear 4; Left Shoulder 5; Right Shoulder 6; Left Elbow 7; Left Wrist 9; Right Wrist 10; Left Hip 11; Right Hip 12; Left Knee 13; Right Knee 14; Left Ankle 15; Right Ankle 16; Left Hip 11; Right Hip 12; Left Shoulder 7; Right Elbow 8; Left Shoulder 9; Right Elbow.



B. Data Generation

Fig. 1. Data Generation process

Each person is cropped in yolo v3 after the image is sent to it. Following cropping, the photos are sent to Net Lightning v3, which is designed for real-time applications, whereas Thunder is designed for applications that need a high degree of accuracy. Prediction. The 17 key points—ear, eye, nose, neck, and so on—that are obtained after passing through it are then stored, plotted by the mat plot library on a black image with the same dimensions as the input, and then saved in.png format. Later, the CNN will be trained using these data.

C. Data description

While video is used as an input, the standard we've discovered for the data we've gathered is around 30 frames per second (FPS). We analyse roughly three frames for every one second of the video, which helps us accomplish a faster classification and real-time deployment.

Each individual player is identified using yoloV3, passed to move net lightning, and the 17 key points are utilised to plot in a black image with the same dimensions as the input, as seen in figure 1.

A total of 5588 pre-processed photos make up our dataset's two classes, of which 1676 files are used for validation and 3912 files are used for training. The 3912 photos are utilised by CNN during training.

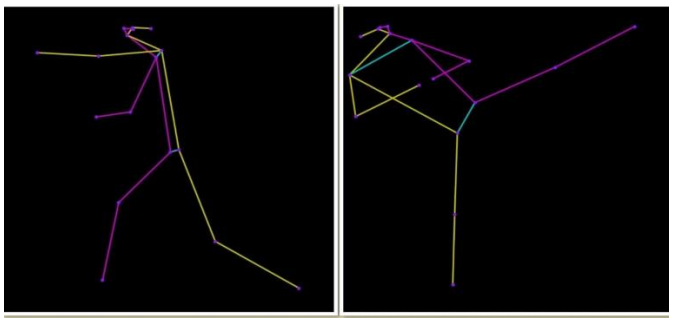


Fig. 2. *Data generated w.r.t 17 key-points*

The total data generated between non-bowling and bowling action is 2985 and 2603. The bowling action is considered for those images where the action of bowling is observed moving one arm above the shoulder level all these images are grouped together forming the database for bowling and then the once with batting, keeping and umpire position are classified as non-bowling actions.

IV. METHODOLOGY

A. Data Flow and transformation

Fig. 3. *Data flow for ROI (*region *of interest)*

Each individual is identified using Yolo V3's video image, and the bounding box coordinates for each frame are then provided to Move Net Lightning, which then locates each individual's key point and passes it to a black image, where it is plotted.

To determine if a person is bowling or not, CNN receives the pre-processed image once it has been collected, rescaled to 180x180, and normalised. These results are provided for each frame together with the bounding box and the original frame for that frame. For presentation, only the bounding box that complies with the demands of the bowling action is taken into account. This outcome is then printed with a bowling motion relative to the original image onto the bordering box. We can more easily compute ROI because each frame is analysed separately and then after the other (region of interest). Since the move net's key-points are inaccurate for the body pose, as shown by the results of the multi-position detection, Figure 3 demonstrates how each individual is recognised and delivered to the move net separately.

B. CNN Architecture

The Rectifier Linear Unit (Re-LU), on which the Convolution Neural Network (CNN) is based, has been found to have time efficiency for both training and testing. Re-LU is a particular kind of activation function that, if the input is positive, produces the input as the output; otherwise, it produces a null value. The input image is scaled down to 180x180 in preparation for CNN processing.

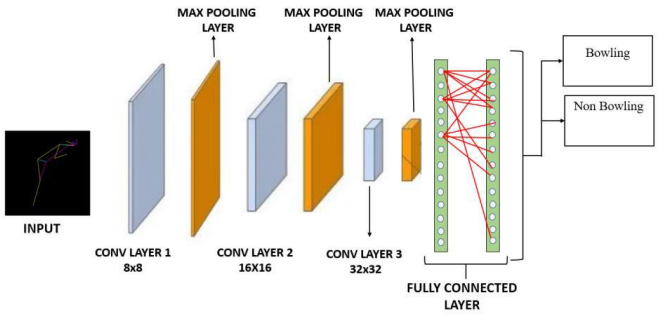


Fig. 4. *CNN architecture*

The CNN is designed to have 3 convolution layers along with 3 max pooling layers and one fully connected or dense layer. There are 2 output classifications violence and non- violence action.

The input to CNN is the picture created by sketching the focal points on a dark image. The CNN is used to categorize an activity as either Violence -related or non-violence-related.

To eliminate false positives, non-violence activity is evaluated if the model's confidence level for a classification of bowling is less than 65%. There are 992,146 total parameters in the network, with an 8x8 convolution layer as the initial convolution layer, a 16x16 convolution layer, and a 32x32 convolution layer following (cony layer 3). There are 991296 parameters in the dense layer. The convolutional neural network in use has the architecture shown in Figure.

RESULT AND DISCUSSION

The system was created to produce an accurate result, however it was discovered that it would need to analyse each frame separately, lengthening the time needed to evaluate each frame. As a result, pipe-lining for real-time deployment is now possible, allowing us to handle the multi-processing task for each individual frame.

Fig. 5. *Model accuracy and loss during training*

The accuracy during training was seen to be 97.34, while the accuracy during validation was 95.19. The validation accuracy is seen to be 95.19 since all bowling poses and actions are consistent across all bowling styles. The validation loss was 0.144 and the training loss was 0.0729; these results were obtained using a batch size of 64 and 10 epochs. The model was halted at the 10th epoch after it was discovered that the accuracy was near in the final three phases. If the target size was adjusted to change for bowling, keeping, and batting, this accuracy may have a different result. Analysis of the data from tournament matches, where a lot of information is gathered, is one of the biggest obstacles.

Fig. 5. *Violent and Nonviolent data analysis*

V. CONCLUSION

In conclusion, a project to identify and respond to violent acts through the use of surveillance cameras offers a potential way to increase public safety. The system is capable of properly detecting violence in real-time thanks to the combination of cutting-edge computer vision, artificial intelligence, and machine learning algorithms.

Future improvements: Multi-camera support, sophisticated deep learning algorithms, and real-time video analytics can all be added to the violence detection system in the future. The precision and dependability of the system can be increased by including other sensors, such audio sensors. Additionally, the usage of block chain technology can improve the system's security and privacy.

VI. REFERENCES

[1] Y. Wang, J. F. Doherty, and R. E. Van Dyck, "Moving object tracking in video," in Proceedings - Applied Imagery Pattern Recognition Workshop, 2000, vol. 2000-January, pp. 95–101.

[2] B. Tian, Q. Yao, Y. Gu, K. Wang, and Y. Li, "Video processing techniques for traffic flow monitoring: A survey," in IEEE Conference on Intelligent Tranhuman actionation Systems, Proceedings, ITSC, 2011, pp. 1103–1108.

[3] J. M. B. Oñate, D. J. M. Chipantasi, and N. D. R. V. Erazo, "Tracking objects using Artificial Neural Networks and wireless connection for robotics," J. Telecommun. Electron. Comput. Eng., vol. 9, no. 1–3, pp. 161–164, 2017.

[4] S. Walker et al., "Systems and methods for localizing, tracking and/or controlling medical in-struments," 2017.

[5] M. Buric, M. Pobar, and M. Ivasic-Kos, "An overview of action recognition in videos," 2017 40th Int. Conv.Inf. Commun. Technol. Electron. Microelectron. MIPRO 2017 - Proc., pp. 1098–1103, 2017.

[6] P. Viola and M. Jones, "Managing work role performance: Challenges for twenty-first century organizationsand their employees.," Rapid Object Detect.

using a Boost. Cascade Simple Fea-tur., 2001.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, 2005.

[8] N. Wang, J. Shi, D. Yeung, J. J.-P. of the IEEE, and undefined 2015, "Understanding and diag-nosing visual tracking systems," openaccess.thecvf.com.

[9] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental compari-son," Pattern Recognition., vol. 76, pp. 323–338, 2018.

[10] Heidari and P. Aarabi, "Real-time object tracking on iPhone," in Lecture Notes

[11] "Multi-Column Deep Neural Networks for Offline Handwritten Chinese Character Classification," by D. Cireşan and U. Meier, in Proceedings of the International Joint Conference on Neural Networks, 2015, vol. 2015-Septe.

[12] Joint CTC-attention based end-to-end speech recognition utilising multi-task learning, S. Kim, T. Hori, S. W.-2017 I. International, and U. 2017, ieeexplore.ieee.org.

[13] Recurrently target-attending track-ing, Z. Cui, S. Xiao, J. Feng, S. Y.-P. of the IEEE, and U. 2016, openaccess.thecvf.com.

[14] "SANet: Structure-Aware Network for Visual Tracking," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2017, vol. 2017-July, pp. 2217–2224. by H. Fan and H. Ling.

[15] "Learning to track at 100 FPS with deep regression net-works," Lecture Notes in Computer Science, vol. 9905 LNCS, 2016, pp. 749–765.

[16] In Proceedings - IEEE International Conference on Circuits and Systems, G. Ning et al., "Spatially supervised recurrent convolutional neural networks for visual object tracking," 2017.

[17] Farhadi, A., and Redmon, J. 2017. Yolo9000: Better, Faster, Stronger, preprint on arXiv.