Suspicious Activity Detection in Recorded and Live Surveillance

Dr. G. Ramesh Chandra
Computer Science and Engineering
VNRVJIET
Hyderabad, India
rameshchandra_g@vnrvjiet.in

Alekhya Lakshmi Mannem
Computer Science and Engineering
VNRVJIET
Hyderabad, India
alekhyamannem0308@gmail.com

Chushmitha Battula
Computer Science and Engineering
VNRVJIET
Hyderabad, India
chushmithabattula99@gmail.com

Naga Shreya Kondepati
Computer Science and Engineering
VNRVJIET
Hyderabad, India
kondepatinagashreya@gmail.com

Sai Shivani Gurram
Computer Science and Engineering
VNRVJIET
Hyderabad, India
shivanigurram2001@gmail.com

***Abstract*— Video surveillance is a tiresome task to perform by a human. Anomaly events are factored by body gestures and head movements of a person. Body gestures of a crowd or a person are only considered for video surveillance when they are in a detectable range from the camera and head movements are considered when a person's face is close enough to the video surveillance, so the necessary facial features are examined.**

In this paper, we aim to detect abnormal activities by first calculating the facial distance to trigger the necessary modules based on the generated focal length through the camera. The abnormal activity detection module is triggered when the threshold distance is crossed and the head movement module is called when the face detected is close enough to the camera. The abnormal event notification is promptly sent to the entitled person so that necessary actions are taken.

Keywords*— *Facial detection, focal length, body gestures, threshold distance, head movement, OpenCV, Keras, mediapipe.

I. INTRODUCTION

Starting with an integrated home security system, surveillance cameras have taken over a variety of settings including public spaces, traffic junctions, and government facilities. The reach of surveillance has increased and so has authorised personnel's inability to constantly monitor criminal and unethical activities along with vandalism.

Automation of video surveillance places a crucial role in keeping public premises safe and secure at all times of the day. Abnormal activities should be detected almost immediately in sectors like banking to detect fraud and provide investigation, healthcare to monitor patients, education to monitor student behaviour and ensure campus security and other events like traffic control, defence systems, and sports venue.

The possible suspicious activity is detected by factoring them into body gestures and head movements. Head movements [1] are a significant radical method of human-computer interaction. It has also received scholars' attention being an effective and simple communication method, [2,3] serving a wide range of monitoring applications. Detecting head movements with multiple landmarks contributes to classifying the suspicious person, in this paper, only six landmark points of the detected face i.e., the corner of the eyes, the nose, the corner of lips, and the chin are considered for abnormal head movements. Head movements detected have a lot of applications in the modern world including [4,5] contactless commands made to the inertial sensors that act upon the instructions of the direction mentioned. To estimate the head movements, the nose is taken as a node, and with the mediapipe, landmarks are coordinated.

Although this seems simple when the face is detected and landmarks are pointed, it is not the vital solution when there are multiple faces detected and the distance is far from the camera. [6] Anxiousness detected from the region of interest (forehead) is only done when the forehead detected from the camera is within the limit of 30 cm which is not possible to detect in public places with heavy crowds.

The distance of the detected objects helps in figuring out the right regions to check for abnormal activity occurrences. Distance is first calculated for the detected face to check whether we need to perform the head movements module or the abnormal activity detection module. The object distance in the frame is calculated from different angles away from the camera [8]. The distance of a moving object from frame to frame is analysed by tracking both the left and right cameras [7],3D construction along all the axis allows for identifying the centroid of an object captured in the camera so that Euclidean distance of the object is calculated.

There are different methods to identify an anomaly event in videos such as CNN for object detection, as labels serve as learning signals in CNN it wouldn't be an efficient method. Using sparse combination learning[9] to detect the anomaly in videos initially removes unwanted data by resizing each frame, extracts spatiotemporal features, and calculates 3D gradient features based on spatial coordinates.

An alternative way is to use spatiotemporal auto encoders[13] which learn the spatial and temporal features. These methods rely solely on unlabelled videos which contain fewer abnormal events, thus achievable in real-world settings.MIRank framework [11] is used to detect the anomaly, normal and abnormal video segments are viewed as instances. To detect anomalous videos, SMG NET[12] is incorporated into the architecture, using the skip connection, randomized abnormal events are added, and the decoder shows variations in the reconstruction done for large datasets.

Along with the above-mentioned papers, this research paper intends to use each of the modules conditionally and make the system simpler and more efficient. If the facial

distance is greater than 200 then the abnormal activity detection module is called, else the head movement module is triggered. The notifications are simultaneously sent to the registered mobile numbers.

The remaining paper covers related works in section II with the details of the proposed system mentioned in section III, the results given in section IV, and the conclusion and future scope in section V.

II. Related Works

Wei et al., [1] used Microsoft Kinect and Kinect for Windows SDK to estimate the head poses. A discrete (HMM) classifier is applied to estimate the action to be a head nod or a shake based on the direction of the head movement. An equal count of head nod, shakes, and other gestures are manually interpreted, and an accuracy of 86% is achieved.

Al-Nuimi et al., [2] implemented a 3D model that calculates the necessary landmarks from the detected face. The nodes are constructed using the face mesh function and landmarks considered are used to find the orientation by taking the extended location between the considered focal points. Low loss with high accuracy is stated in the process.

Sharma et al., [3] intended to find student engagement in their respective online courses through web cameras. concentration index was introduced to estimate the success of the student. The engagement activity is determined by three levels fully, marginally, and not engaged based on their head rotation, facial gestures, and eye rotation.

Atoum et al., [4] presented an automatic online exam proctoring system based on multimedia analytics. Detection of eye gaze, mobile phones etc. are combined with the temporal sliding window to check any fraud done in the test. Extensive results are shown to achieve the accuracy and efficiency of the system.

Carreira et al.,[5] work deals with creating a new dataset called 'kinetics' derived from YouTube videos. The two-stream inflated 3D CNN model is introduced where the filters are inflated. Inflation improves the action classification capability over the datasets HMDB-51 and UCF-101 making them reach a value of 80.2% and 97.9% respectively.

Mossad et al., [6] used an approach to calculate real-time heart rate associated with a face-based video. This method uses Eulerian, Lagrangian transformers and bandpass filters. It can estimate heart rate only when the participants are about 30cm away from the camera.

Ningthoujam Johny Singh et al., [7] used the binocular trigonometry theory to estimate the depth in the image by taking the relative distance location of 2 centroid images captured for the single object from the right and left direction. The stereo vision system showed the minimum error possible

Masoumian Armin et al.,[8] showed how image capturing using 3d cameras is an expensive approach and designed a deep learning framework that contains 2 segments i.e., object detection and depth estimation in an image. Depth estimation is done by using relative distances between the objects which are achieved by using a deep autoencoder network.

Lu, Cewu, et al., [9] used sparse combination learning. They removed unfavourable information by recomputing each frame into patches, extracting spatiotemporal features, and computing 3d gradient features. Only features at the same spatial location are used together for training and testing.

Wang et al., [10] used an S2-VAE combination one being Sc-VAE; a shallow one to filter out clip samples that are normal, and the other being Sf-VAE; a deep network to learn better and identify abnormal events at both local and global levels. The proposed model improved accuracy by 12.2% and at the global level by 2-4% varying on datasets.

Sultani et al., [11] proposed a MIRank framework where video segments are seen as instances, normal and abnormal videos are viewed as bags. The model is based on the fact that anomalous events are rare compared to normal activities. The dataset has 1900 videos with 13 different abnormal activities.

S. Chandrakala et al.,[12] incorporated SMGNet which helped in reviving the abnormal events by not only sticking to the normal patterns since that would solely check the normal events and does not consider events with slight proximity of different patterns, randomized abnormal events are added using the skip connection.

Chong et al., [13] proposed a model that used spatiotemporal architecture which includes spatial feature representation and temporal evolution of the spatial features. This is an unsupervised method and training an autoencoder takes a lot of data and time for processing.

Georgescu et al., [14] work shows that background is not a variant factor in detecting abnormal events. Input is taken with the time series as t - k, t, t + k which further reconstructs upon object detection and skip connection. Each time the classifier takes the absolute differences between these segments and uses gradient descent.

Jianfeng Wang et al., [15] generated a face mask from the highlighted face features which are mapped with the actual face image, feature extraction is done by doing the absolute differences generated. CN is applied to classify these. It is the first paper to show results on the AffectNet dataset.

The above-discussed research papers used body gestures and head movements over the trained images, but they have not actually combined both modules to work conditionally. The head movements have a limited distance w.r.t the camera and the abnormal activity detection solely does not provide the detection since the architectures involved in each work are complex and they require high-end camera resolution to calculate the environment and depth of each object in a frame. This paper uses the advantage of object distance so that the system is made simple. A sequential model is used which helps in achieving the minimum reconstruction error loss and an additional alert feature when an abnormal event has occurred from either of the modules.

III. Proposed System

The system is built by integrating both head movements and abnormal activity detection modules to identify peculiar activities correctly. A user interface is provided and video inputs are given through the web camera or the saved videos. The alerts are sent to the authorized personnel. Video input is passed through the distance module.

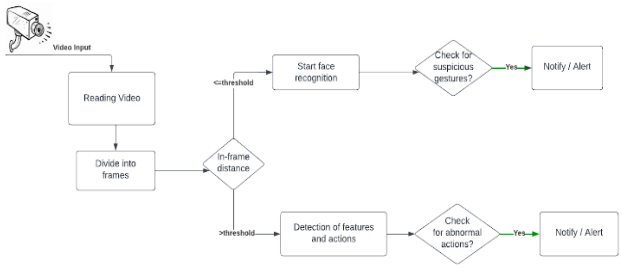


Figure 1: System Architecture

A. Module-1: Distance calculation

Video is sent to the distance module and using the tangent angle relation between the object in the object plane and the image plane, the focal length is calculated, and necessary modules are triggered based on the distance calculated.

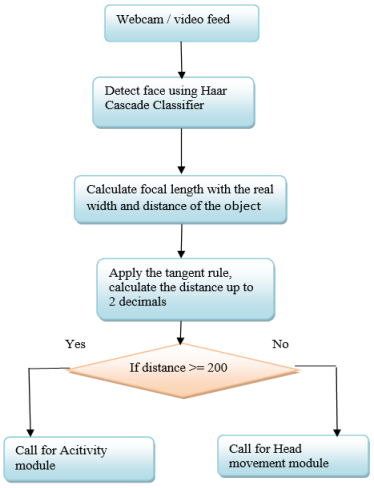


Figure 2: Flow diagram of the distance calculation system

(i) Face detection and generating the focal length

The reference image with its real focal length and width is declared initially. Face width is calculated by generating the width with multiscale facial detection. The tangent angle functionality of the CMOS sensor and object plane is used to identify the focal length.

(ii) Measuring distance in the frame

The frames are read from the video input either through live or saved videos. Distance for the face in each frame is measured by reference to the value of real face width and estimated face width through face detection.

(iii) Module Declaration w.r.t distance calculated

If the distance from the image is measured to be greater than 200 centimetres, the abnormal activity module is triggered for further monitoring; else head movement-based suspicious detection module is triggered.

B. Module-2: Head movement detection

The head movement module is triggered when the distance of the detected face is to be less than 200 cm away from the camera so that the features can be easily distinguished. The algorithm used by this module to detect suspicious activity performed by head movements identified is briefly explained diagrammatically in *Figure 3*.

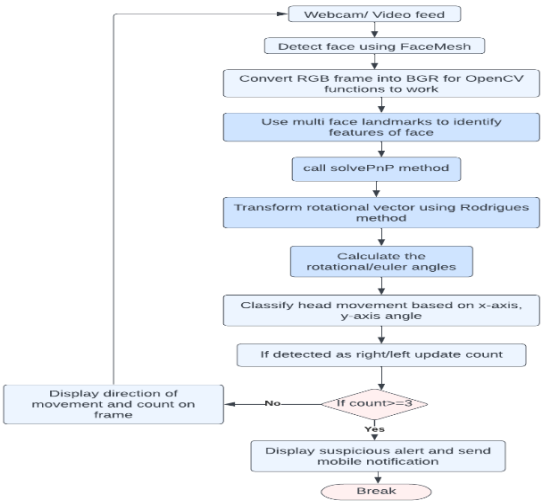


Figure 3: Flow diagram of the head movement detection module

The mentioned instructions below are used to describe the algorithm of the head movement detection process, in-depth for a further detailed understanding.

(i) Initialize the facial landmarks detection model

The face_mesh method from mediapipe. solutions class is used so that a face mesh object is created and face detection is done. min_detection_confidence which shows the model confidence in detecting all landmark positions is set as 0.5.

(ii) Capture and process the image

The video is captured and read (either live or recorded), and for processing the image in mediapipe we need to convert the colour space to RGB format for the landmark detections and convert it back to BRG for the OpenCV library.

(iii) Retrieve landmark locations, and assign the coordinates

The 6 key points from the possible 468 landmark locations are retrieved for head movement detection. The 2D coordinates of the reference matrix of the PnP problem are assigned by calculating the x, and y-axis coordinates of focal points and multiplying them with image width and height respectively. The 3D coordinates for the face model are assigned considering the axis.

(iv) Apply the PnP problem, get the face orientation

Applying all the 2D,3D, camera, and distance matrices in the PnP equation we get rotational and translational vectors. The nose is used as the focal point and the coordinates along with the coordinates of x-axis and y-axis help in identification as looking forward/Down, left/right respectively.

(v) Evaluating directions, notifying the suspicious activity

A 'look' variable is assigned to 0 and it is incremented by one each time when the face is either looking to the left or right directions. The threshold of look to be 3 is considered to be suspicious in this paper since it shows significant consideration for an abnormal activity to occur. Then a notification is sent to registered mobile numbers with 'suspicious alert' as a message by making Twilio API calls.

Algorithm - HeadMovementDetection

Step – 1: *Read input from webcam or as a video file*
- ***cap ← cv2.VideoCapture(0) or cap ← cv2.VideoCapture("video path")***

Step – 2: ***while*** *cap.isOpened()*
- *Read* frame into *image* variable.
- Construct a mesh and store it into the results variable , ***results ← face_mesh.process(image)***
- Initialise empty 3D matrix, 2D matrix as ***face_3d*** *and* ***face_2d*** *for face and reference images respectively.*
- *Consider* ***idx*** *variable present in the* ***face_landmarks*** *dataset and perform a match with the values [33,262,1,61,291,199]*

if true ***then***
(i) *Calculate the* focal length as ***focal_length ← 1*image_width***
(ii) Obtain translational and rotational vectors as output by adding camera matrix and distortion coefficient matrix to the solvePnP method.
(iii) Convert rotational vector into matrix using ***Rodrigues*** method.
(iv) Obtain rotational angles in both x-axis and y-axis from rotational matrix using ***RQDecomp3x3***
(v) Classify head movement as :

if *y< -10* ***then*** *'looking left'*
else if *y>10* ***then*** *'looking right'*
else if *x<-10* ***then*** *'looking down'*
else *'Forward'*

(vi) ***If*** in above case, movement is either *'looking left'* or *'looking right'* ***then***
look ← look+1

end if

- ***if*** *look>=3* ***then***
print *'Suspicious Alert'* text on screen
send ***notification*** to mobile phone using Twilio API method calls.
break

Step -3: Release the cap object using ***cap.release()***

Algorithm 1: Algorithm for head movement detection

C: Module -3: Abnormal Activity Detection Module

The abnormal activity module is triggered when the facial distance calculated is greater than 2 meters. The algorithm used by this module to detect suspicious activity performed by body gestures identified is briefly explained diagrammatically in *Figure 4.* The below-mentioned steps indicate the summary of the mentioned abnormal activity module algorithm.

(i) Preprocessing and Building the model

Conversion of frames into grayscale images is done for the reduction of computation complexity using the standard NTSC formula. A sequential model total of seven layers is added with filters, strides, padding, and activation function 'tanh'. The first two layers are convolutional 3D layers, the next three being convolutional LSTM 2D layers followed by convolutional 3D transpose layers. Adam optimizer is used for faster computation power along with mean squared loss calculation, which is taken as an accuracy metric.

(ii) Classifying into Abnormal Activity using the CNN Model

The model is later trained with the avenue dataset to learn features. To learn more hidden and complex features 5 epochs are used in this paper. After running the epochs parallelly mean squared loss function is used to calculate the loss and

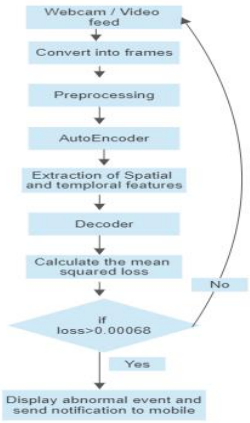


Figure 4: Flow diagram of the abnormal activity detection module

evaluate the model. If the loss is found to be greater than 0.0068 then we declare it as an abnormal event. After the identification of abnormal activity, the notification function is activated. The purpose of this function is to send a text message to saved mobile numbers using Twilio API messaging.

Algorithm – Activity Detection

Step - 1: Read input from webcam or as a video file
- **cap<-cv2.VideoCapture(0) or cap<-cv2.VideoCapture(path)**

Step – 2:**while** cap.isOpened()
- extract videos one by one and convert them into frames folder and store all these frames into images variable
- convert images into gray scale images and resize them to (227,227) with certain interpolation value
- Sequential Model of CNN is used where a total of seven layers are added **two Conv3D layers , three convolutional LSTM 2D layers and two Conv3D transpose layers.**
- Running dataset with 5 epochs parallelly with batch size 1.
- Calculate the mean squared loss:
  - **if** loss>0.00068 **then**

    print 'Abnormal Event Detected' on screen

    send **notification** to mobile using Twilio API method calls

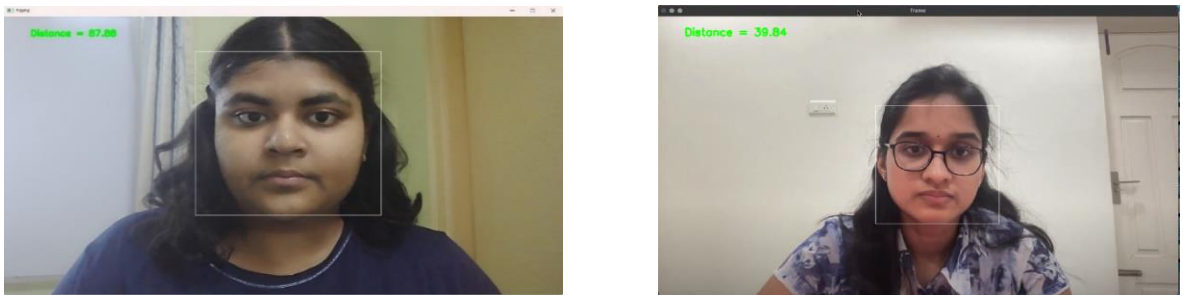
    break

Step – 3:Release the cap object using **cap.release()**

Algorithm 2: Algorithm for abnormal activity detection module

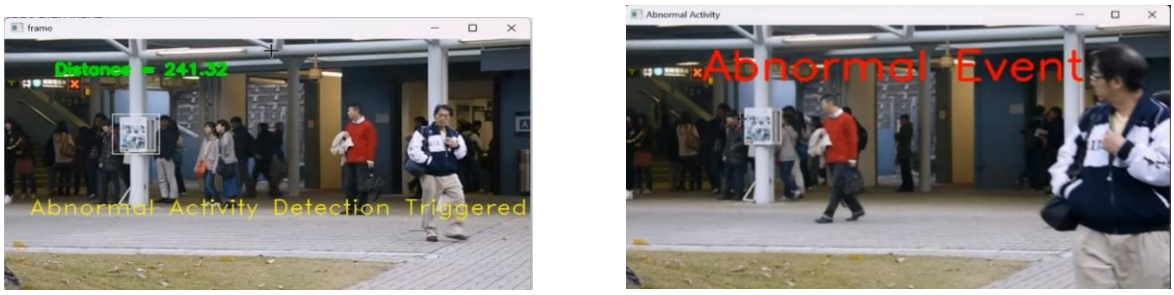
IV. RESULTS

The system with the help of an interactive user interface reads desired input of a user, i.e., either through video or webcam feed. As the system's first module aims at finding the distance from the camera to a person which can be seen as in *Figure 5(a)*– where a recorded video clip is given as input and *Figure 5(b)* –a webcam is used to capture the live feed in real-time. The captured face is shown in the rectangular frame around it. The distance calculated is used as a conditional value, which is compared to a threshold of '2m'. If the calculated value is greater than the threshold then the abnormal activity detection function is triggered as shown in *Figure 6*, where initially the distance for a successful frame is calculated as 241 centimeters as in *Figure 6(a)*, and each frame is processed to detect an abnormal event. The frame where an abnormal event is detected is displayed in *Figure 6(b)*.



(a) *(b)*

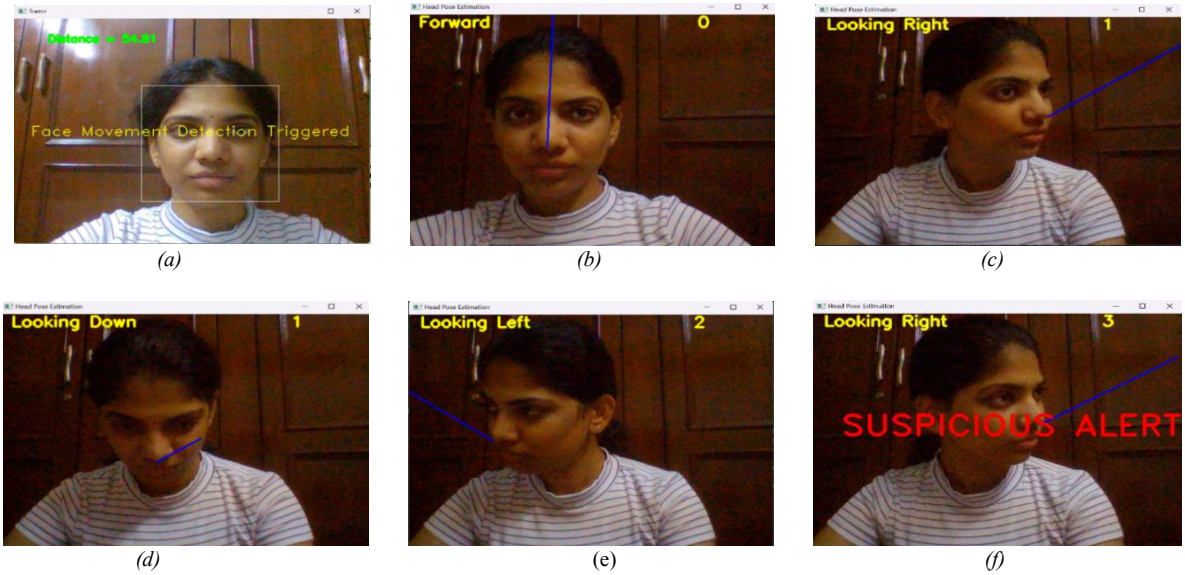
Figure 5: Distance Calculation from Camera to person (a) Using recorded clip as input (b) Using webcam feed as input



(a) *(b)*

Figure 6: Abnormal activity detection function getting triggered (a) Displaying measured distance and module to be triggered (b) Frame where the abnormal event was detected

When the distance calculated is less than the threshold, the head movement detection function will be triggered as shown in *Figure 7(a)*. The function will then work to detect suspicious actions based on head movements being labelled as left and right associated with a counter variable as in *Figure 7(b), 7(c), 7(d), and 7(e)*. When the counter variable attains a value of three i.e., three suspicious head movements have occurred then a suspicious alert is displayed on the screen as in *Figure 7(f)*.



(a) *(b)* *(c)*

(d) (e) *(f)*

Figure 7: Suspicious activity based on head movement (a) Displaying distance calculated and module to trigger (b) Head movement being labelled as 'Forward' (c) Head movement being labelled as 'Right' (d) Head movement being labelled as 'Down'r (e) Head movement being labelled as 'Left' (e) The counter value becomes three and marked as 'suspicious alert'

In the above steps, when the functions to detect abnormal activities and suspicious events based on head movement come into the picture and the deductions are as 'abnormal event' or 'suspicious alert' respectively, then notifications to registered mobile numbers are sent.

V. Conclusion and Future Scope

The system successfully identifies entities based on proximity to the camera to perform distance-dependent functions. These functions focus on identifying abnormal activities when the distance has been greater than the set threshold, otherwise the suspicious activity detection based on head movements is performed. Both of them play a

crucial role in determining abnormal events in any restricted setting of concern. The feature to make it work on recorded videos and webcam feed makes it an efficient system to determine any abnormality and to gather insights based on this suspect. An additional part of the system is the ability to notify concerned mobile numbers when such an event is detected to take timely action.

The system has the scope to be trained on other datasets to work on different surroundings and restricted settings to increase the range of abnormal activities that can be identified. Multiple facial features, expressions, and body language can be used as a basis for suspicious activity detection that could be coupled to the system to monitor an individual more deeply for better deductions. The system as a whole with future additions can directly be integrated with the surveillance cameras and be deployed in surveillance rooms for timely and automated monitoring.

REFERENCES

[1] Wei, Haolin, Patricia Scanlon, Yingbo Li, David S. Monaghan, and Noel E. O'Connor. "Real-time head nod and shake detection for continuous human affect recognition." In 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1-4. IEEE, 2013.

[2] Al-Nuimi, Arqam M., and Ghassan J. Mohammed. "Face Direction Estimation based on Mediapipe Landmarks." In *2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*, pp. 185-190. IEEE, 2021.

[3] Sharma, P., S. Joshi, S. Gautam, S. Maharjan, V. Filipe, and M. J. Reis. "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. arXiv 2019." *arXiv preprint arXiv:1909.12913*.

[4] Atoum, Yousef, Liping Chen, Alex X. Liu, Stephen DH Hsu, and Xiaoming Liu. "Automated online exam proctoring." IEEE Transactions on Multimedia 19, no. 7 (2017): 1609-1624.

[5] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308. 2017.

[6] Ayed, Mossaad Ben, Sabeur Elkosantini, Shaya Abdullah Alshaya, and Mohamed Abid. "Suspicious behavior recognition based on face features." IEEE Access 7 (2019): 149952- 149958.

[7] Singh, Ningthoujam Johny, and Kishorjit Nongmeikapam. "Stereo system based distance calculation of an object in image." In 2019 Fifth International Conference on Image Information Processing (ICIIP), pp. 29-34. IEEE, 2019

[8] Masoumian, Armin, David GF Marei, Saddam Abdulwahab, Julian Cristiano, Domenec Puig, and Hatem A. Rashwan. "Absolute Distance Prediction Based on Deep Learning Object Detection and Monocular Depth Estimation Models." In CCIA, pp. 325-334. 2021.

[9] Lu, Cewu, Jianping Shi, and Jiaya Jia. "Abnormal event detection at 150 fps in matlab." In Proceedings of the IEEE international conference on computer vision, pp. 2720-2727. 2013.

[10] Wang, Tian, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, and Chang Choi. "Generative neural networks for anomaly detection in crowded scenes." IEEE Transactions on Information Forensics and Security 14, no. 5 (2018): 1390-1399.

[11] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6479-6488. 2018.

[12] Chandrakala, S., V. Srinivas, and K. Deepak. "Residual Spatiotemporal Autoencoder with Skip Connected and Memory Guided Network for Detecting Video Anomalies." Neural Processing Letters 53, no. 6 (2021): 4677-4692.

[13] Chong, Yong Shean, and Yong Haur Tay. "Abnormal event detection in videos using spatiotemporal autoencoder." In International symposium on neural networks, pp. 189-196. Springer, Cham, 2017.

[14] Georgescu, Mariana-Iuliana, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. "Anomaly detection in video via self-supervised and multi-task learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12742-12752. 2021.

[15] Chen, Yuedong, Jianfeng Wang, Shikai Chen, Zhongchao Shi, and Jianfei Cai. "Facial motion prior networks for facial expression recognition." In 2019 IEEE Visual Communications and Image Processing (VCIP), pp. 1-4. IEEE, 2019.

[16] Georgescu, Mariana Iuliana, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. "A background-agnostic framework with adversarial training for abnormal event detection in video." IEEE transactions on pattern analysis and machine intelligence 44, no. 9 (2021): 4505-4523.'

[17] Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE signal processing letters 23, no. 10 (2016): 1499-1503f.

[18] Ruiz, Nataniel, Eunji Chong, and James M. Rehg. "Fine-grained head pose estimation without keypoints." In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 2074-2083. 2018.

[19] Cicek, Muratcan, Jinrong Xie, Qiaosong Wang, and Robinson Piramuthu. "Mobile head tracking for ecommerce and beyond." *arXiv preprint arXiv:1812.07143* (2018).

[20] Nyström, Marcus, and Kenneth Holmqvist. "An adaptive algorithm for fixation, saccade, and glissade detection in eye-tracking data." *Behavior research methods* 42, no.1 (2010): 188-204.

[21] Paggio, Patrizia, Manex Agirrezabal, Bart Jongejan, and Costanza Navarretta. "Automatic detection and classification of head movements in face-to-face conversations." In *Proceedings of LREC2020 Workshop" People in language, vision, and the mind"(ONION2020)*, pp. 15-21. 2020

[22] Paggio, Patrizia, Costanza Navarretta, and Bart Jongejan. "Automatic identification of head movements in video-recorded conversations: can words help?." In *Proceedings of the Sixth Workshop on Vision and Language*, pp. 40-42. 2017.

[23] Cao, Zhe, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Realtime multi-person 2d pose estimation using part affinity fields." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291-7299. 2017.

[24] Ambrazaitis, Gilbert, Malin Svensson Lundmark, and David House. "Multimodal levels of prominence: a preliminary analysis of head and eyebrow movements in Swedish news broadcasts." In *Fonetik 2015, Lund, June 8–10, 2015*, vol. 55, pp. 11-16. Centre for Languages and Literature, Lund University, 2015.

[25] Dantone, Matthias, Juergen Gall, Christian Leistner, and Luc Van Gool. "Human pose estimation using body parts dependent joint regressors." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3041-3048. 2013.

[26] Tian, Yuandong, C. Lawrence Zitnick, and Srinivasa G. Narasimhan. "Exploring the Spatial Hierarchy of Mixture Models for Human Pose Estimation." In *ECCV (5)*, pp. 256-269. 2012.

[27] Adjabi, Insaf, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. "Past, present, and future of face recognition: A review." *Electronics* 9, no. 8 (2020): 1188.

[28] Wang, Yitong, Xing Ji, Zheng Zhou, Hao Wang, and Zhifeng Li. "Detecting faces using region-based fully convolutional networks." *arXiv preprint arXiv:1709.05256* (2017).

[29] Vignesh, Kothapalli, Gaurav Yadav, and Amit Sethi. "Abnormal event detection on BMTT-PETS 2017 surveillance challenge." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 36-43. 2017.

[30] Ibrahim, Muhammad Junaid, Jaweria Kainat, Hussain AlSalman, Syed Sajid Ullah, Suheer Al-Hadhrami, and Saddam Hussain. "An Effective Approach for Human Activity Classification Using Feature Fusion and Machine Learning Methods." Applied Bionics and Biomechanics 2022 (2022).

[31] Xu, Yuanyuan, Wan Yan, Genke Yang, Jiliang Luo, Tao Li, and Jianan He. "CenterFace: joint face detection and alignment using face as point." Scientific Programming 2020 (2020).

[32] Minaee, Shervin, Mehdi Minaei, and Amirali Abdolrashidi. "Deep-emotion: Facial expression recognition using attentional convolutional network." Sensors 21, no. 9 (2021): 3046.