Integrated and Deep Learning–Based Social Surveillance System: a Novel Approach

Submitted: 6th April 2022; accepted: 3rd May 2022

Ratnesh Litoriya, Dev Ramchandani, Dhruvansh Moyal, Dhruv Bothra

DOI: 10.14313/JAMRIS/3-2022/22

Abstract:
In industry and research, big data applications are gaining a lot of traction and space. Surveillance videos contribute significantly to big unlabelled data. The aim of visual surveillance is to understand and determine object behavior. It includes static and moving object detection, as well as video tracking to comprehend scene events. Object detection algorithms may be used to identify items in any video scene. Any video surveillance system faces a significant challenge in detecting moving objects and differentiating between objects with same shapes or features. The primary goal of this work is to provide an integrated framework for quick overview of video analysis utilizing deep learning algorithms to detect suspicious activity. In greater applications, the detection method is utilized to determine the region where items are available and the form of objects in each frame. This video analysis also aids in the attainment of security. Security may be characterized in a variety of ways, such as identifying theft or violation of covid protocols. The obtained results are encouraging and superior to existing solutions with 97% accuracy.

Keywords: *Video Surveillance, object detection, object tracking, YOLO v4 algorithm, OpenCV*

1. Introduction

In this day and age, people have begun to rely more on technologies that are called smart, meaning that they can operate and learn on their own at the command of humans and do the needed duties. We have achieved this new step with the assistance of computer vision and deep learning. In planning, operation, and sustainability of contemporary industrial and urban areas, video surveillance is a major factor. The efficiency, safety, security, and optimality of the region, infrastructures, persons, operations, and activities are all aided by video surveillance [1]. Autonomous equipment, cyber-physical systems, and energy-efficient architectures are becoming more common in industrial settings. With the rising use of multi-level structures and greater traffic, pedestrian, and crowd movements, urban landscapes are becoming more densely inhabited. In both industrial and urban contexts, this vertical and horizontal growth of asset and area utilization has led to considerable growth in the implementation of closed-circuit television (CCTV) camera systems to ensure the safety of humans or assets and surveillance of activities. We have begun to perform analysis on photos which was previously confined to textual data. However, analyzing static images was no longer sufficient; we now need to examine videos using an approach similar to that used for still images. Giving a live video feed a real-time output is a new difficulty in the industry.

With the use of CCTV, one can watch an area 24/7, or the video may be retrieved when needed provided it is kept in a secure location. It can be used to prevent crime and assist law enforcement in identifying and solving crimes. With the help of YOLO, a new approach opens to perform detection with an extremely quick architecture, real-time image computation, open-source computer vision (OpenCV) software, and a powerful library of image processing tools.

In our study, we have taken up that challenge and attempted to apply it as much as possible to the field of surveillance in order to tackle the challenges of the modern world. We've been looking for abnormalities in real-time footage. In our research, we selected three real-world anomalies to identify: violations of social distance norms, face mask usage monitoring, and theft detection to determine if a valuable object is being taken in real time and from where in the frame.

One of the main worries of today's citizens is security. In our system, we've combined three functionalities that have proven to be important. In order to try and address some problems of the new common world, we built a working prototype of a multi-purpose smart surveillance system that can be used to detect multiple anomalies in real time and can be used anywhere, from a crowded public space to a quiet museum with a reasonable number of people in the room. The proposed system is less complex and more accurate, as compared to the existing solutions.

This article also includes a literature survey with information about past research in this field. Methodology includes the implementation phase of this research, with the diagrammatic approach used in the Result and Discussion section, which includes all the outcomes drawn by the research. Our conclusion includes a summary of the research made.

2. Literature Review

Our literature search for intelligent systems knowledge, relatable or preceding working prototypes, and algorithms for human and feature recognition was conducted, and the various papers mentioned were found to be very helpful in understanding the problems and the solution of those problems in the context of algorithms and the proposed system. These works appeared publications such as IEEE, Springer, Elsevier, Willey, and Cambridge Press, to mention a few. Papers based on deep neural networks, computer networks, the YOLO algorithm, social distance identification, face mask detection, and video analysis were the focus of our search. In the field of video analysis and object detection, many researchers have done a substantial amount of work and shared their findings with the world. Some of these are discussed below.

Redmon et al. [2] proposed a system using the YOLO algorithm. Classifiers developed for object detection are repurposed for per-form detection. Their design was lightning quick: at 45 frames per second, their YOLO model analyzes pictures in real time. You just need to glance at an image once to guess what things are present and where they are with YOLO. YOLO is a simple and quick algorithm that uses complete photos to train the detection process. It doesn't employ a complicated pipeline. It uses a neural network to anticipate detections on a fresh picture at test time, which allowed us to handle live streams.

The model was simple to build and can be trained on complete photos. YOLO is the fastest object classification detector in the literature, and it pushes the boundaries of real-time object detection. YOLO is also adaptable to new domains, making it excellent for applications that need quick and reliable object recognition.

Suwarna Gothane [3] says in his research that they can quickly detect and identify the items of our attention when looking at photographs or movies. Object detection training is the process of passing this knowledge to computers. The YOLO model is quite precise and can recognise the things in the picture. YOLO takes an entirely new approach. Instead of focusing on individual regions, it employs a neural network to predict anchor boxes and their probability throughout the whole image [4]. YOLO uses a single deep neural network that divides the input image into grid size. Unlike image classification or face detection, each grid size in the YOLO algorithm has a related matrix in the output that informs us if a result is found in that grid size and class of that object.

Gupta and Devi [5] made a study of the YOLOv2 method that is proposed for the identification of objects in pictures with geolocation and video recordings. The primary goal of this study is to detect things in real time, that is, live identification, utilizing a camera and video recordings. COCO, a dataset with 80 classes, was employed in this work [6]. Using the YOLOv2 model, it is simple to recognize items with grids and boundary prediction, and it also aids in predicting exceptionally small things or objects that are far away in the image. Darknet makes it simpler to recognize moving objects in video recordings, and it generates .avi files containing detections.

Jayashri et al. [7] suggest a real-time system for monitoring social distance and avoiding congestion by employing the concept of suggested critical social density. The team is committed to providing innovative, strategic advancements that protect persons and networks. Under the present circumstances of the COVID-19 pandemic, this work has practical value. The pipeline is capable of detecting persons with, without, and incorrectly wearing coverings with reasonable accuracy.

Gupta et al. [8] proposes a simpler way to accomplish this goal by utilizing some fundamental deep learning tools such as TensorFlow, Keras, and OpenCV. The suggested technology successfully recognizes the face in the image/video stream and determines whether or not it is wearing a mask. It can recognize a face and a mask in motion as a surveillance task performance. Optimal parameter settings need to be determined for the CNN model in order to identify the existence of masks accurately.

Kakadiya et al. [9] suggested a system using deep learning to establish a smart camera that observes bank activities and can identify any suspicious conduct. Criminals can be followed based on mobility and weapon presence. The SmartCam immediately transmits a message to the safety committee if any suspicious weapons or actions are observed. The message specifies the sort of warning that has been issued, as well as the type of weapon and number of weapons identified, as well as a web link to a live picture that may be viewed by security personnel.

Patil et al. [10] proposed that theft is among the most widespread criminal activities, and it is on the ascent. It has become one of the world's never-ending issues. Their study used a sophisticated algorithm like CNN which provides an advantage over more standard algorithms such as RNN, SVM, and others. This program correctly recognizes emotional expressions with a higher percentage of accuracy. The Keras toolbox is used for this. They arrived at the above-mentioned makeshift model after experimenting with various layer combinations and iterations.

Chandan G. et al. [11] suggest an approach using the SSD method. In real-time applications, the SSD method is used to identify objects. SSD has also demonstrated outcomes with a high level of confidence. The main goal of the SSD method is to recognize and track numerous objects in a real-time video stream. This model performed admirably on the object trained on in terms of detection and tracking, and it may be used in certain circumstances to identify, track, and respond to specifically targeted objects in video surveillance.

Kumar et al. [12] suggest that for traffic and surveillance applications, object detection algorithms such as You Only Look Once (YOLOv3 and YOLOv4) be used. An input layer with at least one hidden layer and an output layer make up a neural network. Multiple object detection in surveillance cameras is a difficult task that is influenced by the density of items in the monitoring area or on the road, as well as timings. The multiple object detection technique implemented in this work is useful for traffic and various surveillance applications. The dataset is made up of pictures and videos with different levels of light. The system efficiently recognizes several items with high accuracy, according to the results.

Bochkovskiy et al. [13] stated in their research that there are a slew of factors that are thought to increase the accuracy of convolutional neural networks (CNNs). Theoretical explanation of the conclusion, as well as experimental evaluation of combinations of such characteristics on large datasets, is required. Some characteristics, such as batch normalization and residual connections, are appropriate for most models, tasks, and datasets, while others are only suitable to particular models and issues, or only for small-scale datasets. Weighted-residual-connections (WRC), cross-stage-partial-connections (CSP), cross mini-batch normalization (CmBN), self-adversarial-training (SAT), and Mish-activation are all assumed to be universal properties.

As can be seen from the preceding literature study, our fellow researchers saw the need and worked diligently to meet it. Some compared a variety of algorithms in order to arrive at a speedier answer, while others attempted novel techniques to get better results. By reading these, one can get a solid sense of how things work when it comes to recognizing abnormalities in real time and comprehending the obstacles that come with it. Despite having amazing methodologies, observations, and findings, reading the above work reveals that none of our predecessors worked with or attempted to merge multiple-use cases to create a technology that would be multifunctional and adaptable to the demands of the modern world. In our research, we attempted to achieve exactly that, creating a multifunctional smart CCTV to meet today's demands by analyzing past practices, and developing our own practices that work well together to become an effective method with real-time output.

3. Methodology

The suggested research is built using Python 3, OpenCV, and the flask framework of Python. Using OpenCV's machine learning methods, we can educate the machine to discern between diverse user use cases and unauthorized offenders' unique undesired conduct, allowing us to take appropriate action based on the context. We are able to effectively use logic to execute the artificial intelligence idea at hand to recognize and classify the events that occur using image processing strategies and mathematical deductions. Additionally, the system is capable of taking action in response to the current occurrence [14].

The primary goal of this system is to analyze collected video footage for human detection and then further analyze it for any anomalies. An anomaly can be anything the user sets the software to detect. A breach of social distancing laws, a person not wearing a mask, or the detection of theft on video footage could all be considered anomalies for this research.

The procedure begins by scanning each frame of a video stream one by one. This is depicted in Fig. 1, and is also depicted in the block diagram showing the entire sequence of actions.

The object detection framework is the most essential aspect of this research. This is due to the study's

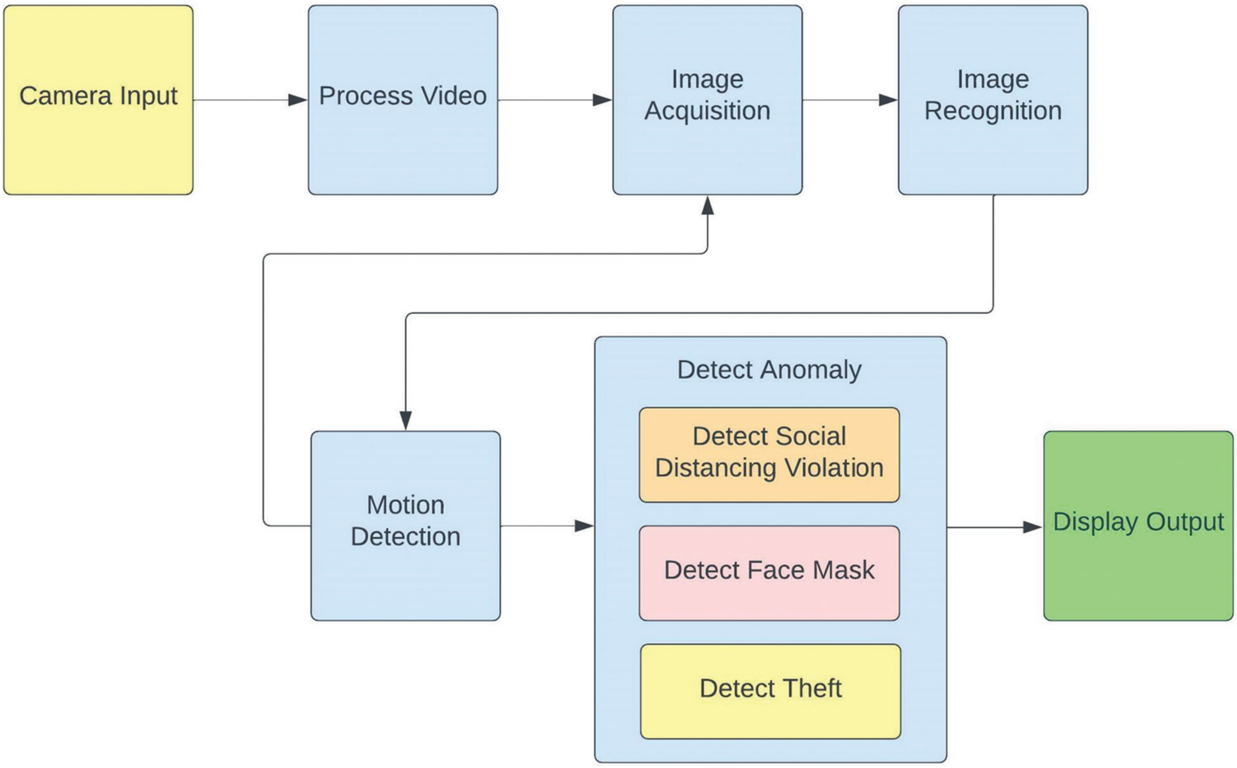

Fig. 1. Block diagram of methodology

component that focuses on establishing a person's position from the input frame. As a result, selecting the most appropriate object detection model is critical in order to prevent any issues with recognizing people [15].

Monitoring: As soon as the system is turned on, its initial inclination is to scan its surroundings for any movement that could occur in the situation under consideration. The primary goal of motion analysis is to minimise large duplicate activity storage. The recording of the movement begins as soon as the camera detects any unknown creature approaching the target.

Masking Frame: Masking is an image processing technique in which a tiny picture fragment is defined and used to affect a bigger image. Setting part of the pixel values of an image to zero and another backdrop value is known as masking [16]. The picture will be isolated. For example, a video is a collection of images that are played in a specific order over a period of time. To build ROI for each frame of the input frame, the OpenCV masking approach will be employed in this study.

Motion Detection: The surveillance system stays silently watching until it detects an unknown creature coming, at which point the camera begins recording the scene in question, which includes a clear view of the item.

3.1. Detecting Anomaly

3.1.1. Theft Detection

After the motion is detected, we save the frame just before the motion, and the next static frame after the motion is also taken. Now both the frames are converted to greyscale and blurred to make a comparison. The photos are converted to greyscale since less information is required for each pixel. Converting to greyscale also divides the luminance and chrominance planes. Luminance is more crucial for identifying visual characteristics in a picture. The image will be blurred: the picture is blurred by using a low-pass filter kernel to convolve it. This can be used to reduce noise. We'll achieve it with the picture blurring averaging approach, which involves convolving an image with a normalized box filter. It simply replaces the core element with the average of all pixels under the kernel region. Then we calculate the image similarity score using the structural similarity function of skimage. The mean squared error (MSE) is a straightforward way to compare photos, but it isn't a good indicator of perceived resemblance. By taking texture into consideration, structural similarity functions seek to remedy this problem. If the similarity score is more than the desired threshold, then we classify that nothing is stolen. On the other hand, if the similarity score is less than the desired threshold then it suggests that the two frames have structured dissimilarity and something is stolen or missing. Furthermore, in the second case, we again use the grey-scale images, apply thresholding to them, and construct a rectangular box where the dissimilarity exists.

3.1.2. Face Mask Detection

If the input is a video stream, the picture or a frame of the video is initially delivered to the default face detection module for detection of human faces. This is accomplished by first enlarging the picture or video frame, then identifying the blob inside it [17]. The face detector model receives this identified blob and outputs just the cropped human face without the backdrop. This face is used as model input that we previously trained. This determines whether or not a mask is present. [18].

To implement the face mask detection function, we will utilize convolution neural networks to train our model, with one exception: we will skip the convolution layer of the feature map and replace it with MobileNets. MobileNets are low-latency, low-power models that have been parameterized to match the resource restrictions of various use cases. So, after converting the input picture to an array, we'll send it to MobileNets. Furthermore, we will perform max pooling, which is a pooling procedure that determines the maximum value for patches of a feature map and utilizes that value to produce a down-sampled (pooled) feature map. We'll next flatten it to create a completely linked layer that we'll utilize to generate output. We chose MobileNets since they have been shown to be quicker than traditional convolutional neural networks in terms of processing speed and parameter use. Although MobileNets appears to be a good option, it has its own drawbacks. They are occasionally less accurate, but for our purposes of creating a model to utilize in real time, they have proven to be more effective because we preferred speed over minor accuracy.

3.1.3. Social Distancing Detection

To detect humans in the frame in this module, we employed the YOLOv4 method, which is an object detection system that is a development of the YOLOv3 model. It is twice as quick as EfficientDet and has comparable performance, thus it is a good fit for us to fulfill our job and provide real-time output. YOLO is an acronym that means "You Only Look Once." Because of its simplified construction, it operates much quicker than RCNN. It's taught to conduct classification and bounding box regression at the same time, unlike quicker RCNN. As a result, we utilize it to find persons in our model. We save all the instances of the output in a set after detecting individuals and bound each person in a rectangular box to determine the centroid of each person later.

Find the centroid of the person detected on the frame using the below formula.

Centroid of rectangle:($(x_1 + x_2) / 2, (y_1 + y_2) / 2$)

where,x-Center = (x_1+x_2) /

2y-Center = $(y_1 + y_2) / 2$

Moving on after finding the centroid, compute the pairwise Euclidian distances between all detected people.

Euclidian distance: $d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$

where,

- (x_1, y_1) are the coordinates of one point.
- (x_2, y_2) are the coordinates of the other point.
- d is the distance between (x_1, y_1) and (x_2, y_2).

Based on these measurements, determine whether any two persons are fewer than N pixels away, which is the accepted threshold pixel distance. The minimum pixel distance varies depending on the camera's height and angle. The social distance protocols are broken if the distance between two centroids is smaller than the threshold distance, and vice versa.

4. Results and Discussion

4.1 Theft Detection

This method is used to find whether an object is moved from its place. If the object is moved or is in motion, the next static frame after motion will be taken to check whether the object is moved or not. If the object goes missing in the static frame after motion then a box will be made on the initial frame taken for comparison. The full procedure is depicted in Figure 2. Figure 2 shows an object resting on the table with no motion detected at this time. However, when the object is moved, as shown in Figure 3, motion is detected, and our system provides us with information about the object being moved as well as from where the object is moved, as shown in Figure 4.

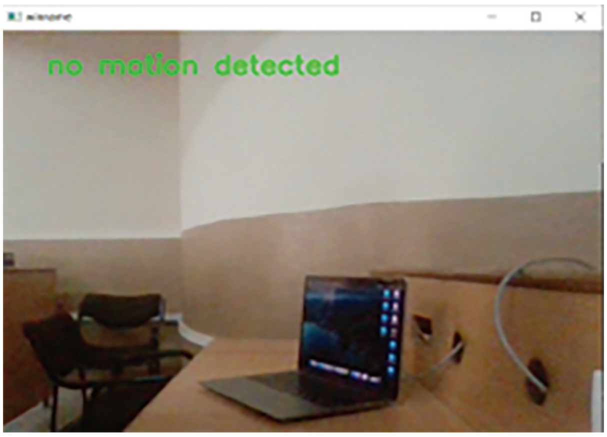

Fig. 2. Frame with object

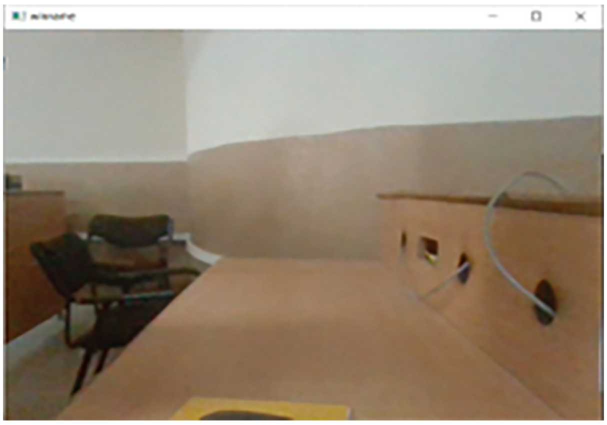

Fig. 3. Frame without object

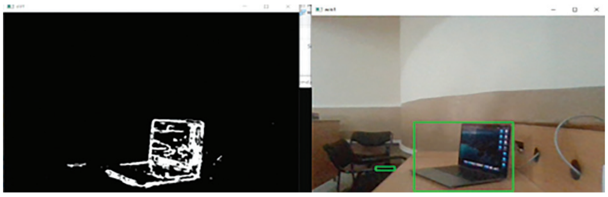

Fig. 4. Missing object highlighted in green boundary and grey-scale image of missing object

Previously, as mentioned in the literature review of a paper by Patil et al. that proposed a solution to the problem of theft detection in which they drew a link between the object and the owner, in order to identify theft if the link between them increased. Even though it is a reliable and effective method, it does not identify the position of the missing object, as our proposed approach does. Also, the previous solution was designed with the goal of detecting theft at airports or similar locations, whereas our proposed solution is better suited to places like bank lockers and museums, where objects of interest are at rest and untouched for the majority of the time, as it would be easier to detect motion and detect missing objects in those settings.

In our theft detection function, motion was identified 100% of the time in decent lighting, but only 91 percent of the time in low or dim illumination. Each time an object was taken or moved in the frame, it was detected with 100% accuracy and the precise location of the missing object was noted. Although the software's recognition of the object's shape was not flawless, it was able to accurately indicate the shape of the missing object more frequently than not.

4.2 Face Mask Detection

In the above Figs. 5 and 6, a bounding box is drawn around the ROI enabling a check to see whether the person is wearing a mask or not. The green colour bounding box depicts the person is wearing a mask and the red bounding box depicts a person is without

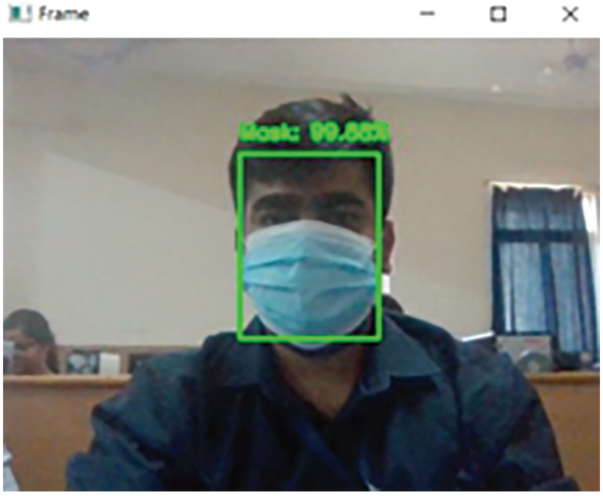

Fig. 5. Person with mask on

mask. Fig. 5 shows that when a person successfully wears a mask, our system correctly predicts that the person is wearing the mask with an accuracy of 99.88 percent, but when the person is not wearing the mask, it predicts that accurately with 100% accuracy, as shown in Fig. 6.

This method works on a group of people in the frame as well, as we can see in Figs. 7 and 8. Accordingly, ROI is created around each face and bounding boxes are assigned. The green box will only be assigned if a person is wearing the mask correctly. In all other cases the red bounding box is assigned. Fig. 7 shows that when a person is not wearing a mask properly, it is accurately recognized as such with a high degree of accuracy, whereas Fig. 8 shows that our system can recognize faces from a fair distance and correctly classify several persons wearing masks or not.

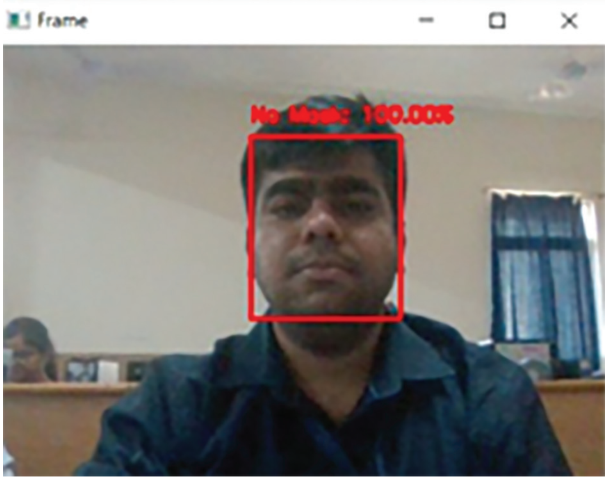

Fig. 6. Person without mask

In contrast to S. Gupta et al.'s work, which used computer vision to detect face masks, we chose to use deep neural networks to detect face masks because a deeper network can learn a more complex, non-linear function, which improves performance. This allows the networks to discriminate between different classes more easily if they have enough training data. In comparison to a network with regular convolutions of the same depth in the nets, we used MobileNets because it significantly reduces the number of parameters. As a result, lightweight deep neural networks are created. Two procedures are used to create a depth wise separable convolution.

We have exhibited the results of our trials, as well as a table of observations (Table 1) that we used to calculate the accuracy of our system. When a human face was an acceptable distance from the camera, the face mask detection algorithm was able to detect a human face with an accuracy of 97 percent and determine whether the person was wearing a mask or not based on the features that could be seen. After repeated tests, the mask identified and undetected accuracy ranged from 92 percent to 100 percent. The optimal setting to run the system was bright light, since it produced the best accuracy of up to 100 percent, but when the light was dim, the accuracy dropped and usually stayed between 86 percent and 95 percent. Overall, the technology correctly predicted whether

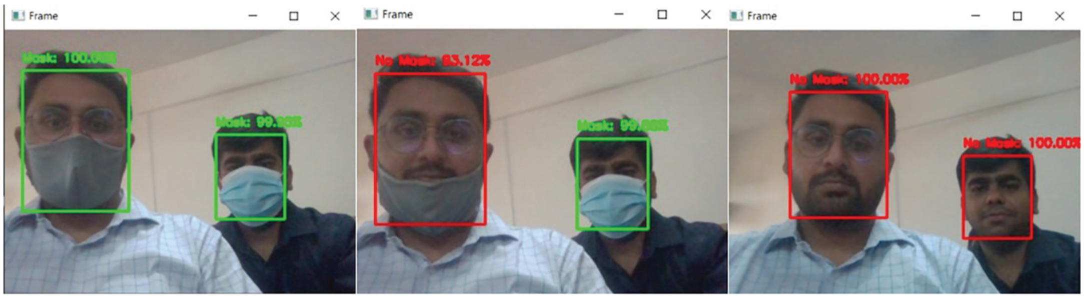

Fig. 7. Two people with and without mask

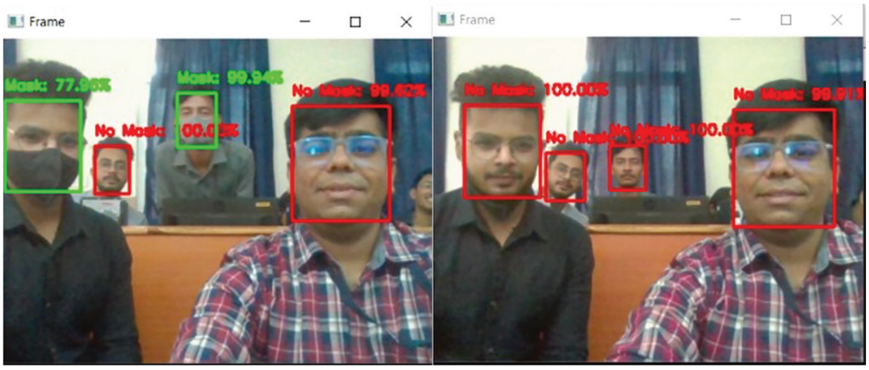

Fig. 8. Group of people with and without masks

Tab 1: Outcomes of face mask detection method

Sr.no.	People in Frame	Lighting Conditions	Face Detected	Correct prediction
1	1	Standard lighting	1	1
2	1	Dim lighting	1	1
3	2	Standard lighting	2	2
4	2	Dim lighting	2	2
5	3	Standard lighting	3	3
6	3	Dim lighting	3	3
7	5	Standard lighting	5	5
8	5	Dim lighting	5	5
9	7	Standard lighting	7	7
10	7	Dim lighting	6	6
11	8	Standard lighting	8	8
12	8	Dim lighting	8	8
13	10	Standard lighting	10	10
14	10	Dim lighting	9	9
15	11	Standard lighting	11	11
16	11	Dim lighting	11	10
17	12	Standard lighting	12	12
18	12	Dim lighting	11	11
19	15	Standard lighting	15	15
20	15	Dim lighting	14	13

or not people were wearing masks 96 percent of the time.

4.3. Social Distancing Detection

As seen in Fig. 9, our system was able to recognize human figures in video frames rather effectively, even when given a frame of crowded individuals, and assess whether humans are obeying or breaching social distance regulations. The experiment was conducted by positioning a camera from a top-down perspective; the observations table is shown as Table 2. Our system was able to recognize practically all human figures from a decent height with a 98 percent efficiency. We have seen that its efficiency only drops by 2 percent, to 96 percent, when the same camera is put at the same angle but under different environmental conditions—in this example, when the lights are dim. We also put our surveillance system to the test against shadows, which might look human-like from afar. In the case of shadows, our algorithm performed admirably, with a 97 percent accuracy in identifying between a genuine human and a shadow. Even though our system is able to detect humans effectively when the camera is not at a top-down approach and is directly in front of humans, as shown in Fig. 10, we can see that it is not able to recognize the distance correctly because it does not factor in the depth, and thus gives us bad predictions. As a result, it is recommended that the camera be placed high in order to obtain better and more precise findings.

The rapid spread of coronavirus leaves the major population vulnerable to getting infected.

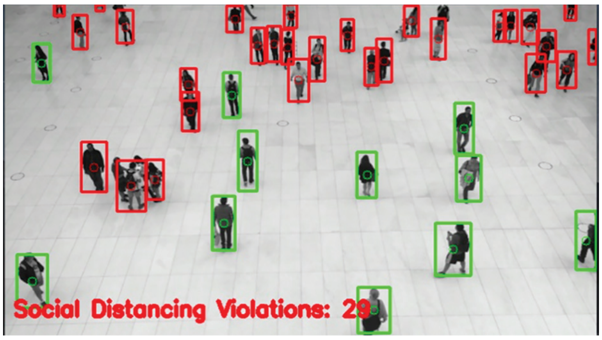

Fig. 9. Predicted simulated results of SSD Inception V2 for 3 classes of chest pain facial expression detection

The preventive health care team along with the technology specialists must remain vigilant and focus on strategic areas. Social distancing is a proven method used to control the spread of any contagious diseases. As the name suggests, social distancing implies that people should physically distance themselves from one another, reducing close contact, and thereby reducing the spread of a contagious disease. The presented research focuses on assuring safe distance among people while at the same time providing the ability to detect theft. This integrated solution works efficiently and accurately. This contribution utilizes advanced image processing concepts along with efficient computing algorithms to solve the issues of social distancing and theft detection and to prevent the spread of pandemics.

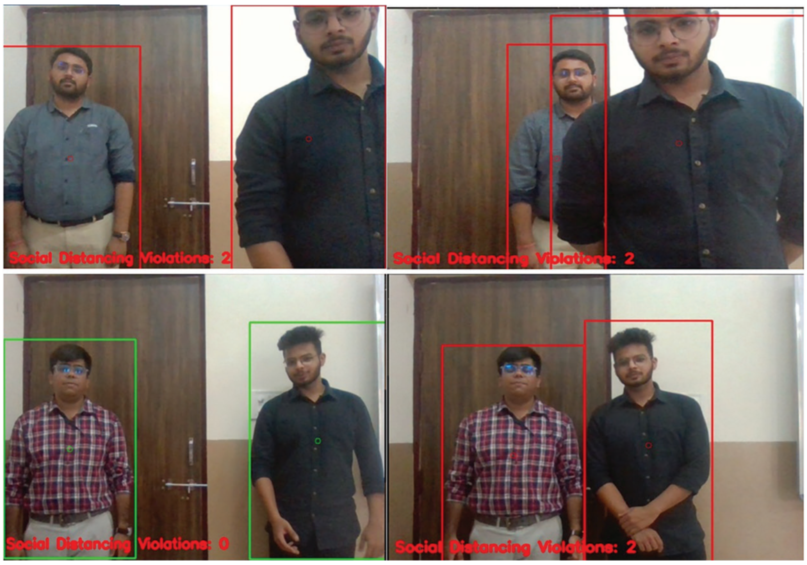
Social Distancing Violations: 2
Social Distancing Violations: 2
Social Distancing Violations: 0
Social Distancing Violations: 2

Fig. 10. Social distancing module used at a non-inclined angle

Tab 2: Outcome for social distancing detection method

Sr.no.	People in Frame	Lighting Conditions	Humans Detected	Correct Prediction
1	10	Standard lighting	10	10
2	10	Dim lighting	10	10
3	12	Standard lighting	12	12
4	12	Dim lighting	12	12
5	18	Standard lighting	18	18
6	18	Dim lighting	18	18
7	22	Standard lighting	22	22
8	22	Dim lighting	22	21
9	29	Standard lighting	29	29
10	29	Dim lighting	29	29
11	32	Standard lighting	32	32
12	32	Dim lighting	31	30
13	35	Standard lighting	34	33
14	35	Dim lighting	33	33
15	40	Standard lighting	40	40
16	40	Dim lighting	37	37
17	44	Standard lighting	42	40
18	44	Dim lighting	43	43
19	50	Standard lighting	50	49
20	50	Dim lighting	48	48

5. Conclusion

One of the most significant precautions in avoiding physical contact that could contribute to the spread of coronavirus is social distancing. Viral transmission rates will be increased as a result of noncompliance with these rules. To implement the proposed features that are crucial to stop the spread of coronavirus—social distancing and wearing face masks—a system was created using Python and the OpenCV library. In the first and second features, we also employed the YOLO method to recognize humans and classify faces, respectively. The current research examines whether or not people were wearing face masks. Real-time video streams and images were used to test the models. The model's optimization is a continual process, and we're fine-tuning the hyperparameters to provide a very accurate answer. Because of its high precision and low error rate, the proposed method can be easily implemented in real scenarios, such as schools, public places like airports, bus stations, banks, tourist attractions, museums, and many more. However, it does not work as well for all camera angles or setups. A top-down camera angle is the best-advised camera angle. We have addressed a need for more intelligent surveillance technologies to be employed in public and private spaces in order to make it easier for persons monitoring those areas to identify abnormalities more quickly and accurately in this work. Then, in a functional prototype of a system, we presented a solution to the problem. For the sake of our research, we categorized different actions such as violation of covid protocols and theft as abnormalities, and we then attempted to develop a multipurpose intelligent surveillance system for the same that has several modes and can work in any mode dependent on the user requirements. In the future, we intend to test our concept in a variety of industries, including service, commerce, and security.

AUTHORS

Ratnesh Litoriya* – Computer Science Engineering Dept. Medi-Caps University, Indore, India, Email: litoriya.ratnesh@gmail.com.

Dev Ramchandani – Computer Science Engineering Dept. Medi-Caps University, Indore, India.

Dhruvansh Moyal – Computer Science Engineering Dept. Medi-Caps University, Indore, India.

Dhruv Bothra – Computer Science Engineering Dept. Medi-Caps University, Indore, India.

*Corresponding author

REFERENCES

[1] H. Liu, S. Chen and N. Kubota, "Intelligent Video System and Analytics: A Survey," *IEEE Transactions on Industrial Informatics,* vol. 9, no. 3, 2013, pp. 1222-1233.

[2] J. Redmon, S. Divvala, R. Grishick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* 2016, pp. 779-788.

[3] D. S. Gothane, "A Practice for Object Detection Using YOLO Algorithm," *International Journal of Science Research in Computer Science, Engineering and Information Technology,* vol. 7, no. 2, 2021, pp. 268-272.

[4] B. Qiang, R. Chen, M. Zhou, Y. Pang, Y. Zhai and M. Yang, "Convolutional Neural Networks-Based Object Detection Algorithm by Jointing Semantic," *Segmentation for Images, Sensors,* 2020.

[5] S. Gupta and D. T. U. Devi, "YOLOv2 Based Real Time Object Detection," *International Journal of Computer Science Trends and Technology,* vol. 8, no. 3, 2020.

[6] G. S., "Real-Time Object Detection with Yolo," *proceedings of the International Journal of Engineering and Advanced Technology (IJEAT),* 2019.

[7] T. K. M, V. P. M., Y. B., J. S. and L. Dr. K., "Video Analytics on Social Distancing and Detecting Mask - A detailed Analysis," *International Journal of Advanced Engineering Research and Science (IJAERS),* vol. 8, no. 5, 2021.

[8] S. Gupta, V. Dhok, A. Chandrayan and S. Tiwari, "Facemask Detection using OpenCv," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 10, no. 6, 2021.

[9] R. Kakadiya, R. Lemos, S. Mangalan, M. Pillai and S. Nikam, "AI Based Automatic Robbery/Theft Detection using Smart Surveillance in Banks," *Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology,* 2019.

[10] S. Patil, M. Shidore, T. Prabhu, S. Yenare and V. Somkuwar, "Theft detection using computer vision," *International Journal of Advance Research, Ideas and Innovations in Technology,* vol. 5, no. 1, 2019, pp. 567-569.

[11] C. G, A. Jain, H. Jain and M. , "Real Time Object Detection and Tracking Using Deep Learning and OpenCV," *Proceedings of the International Conference on Inventive Research in Computing Applications,* 2018.

[12] C. Kumar B, P. R and Mohana, "YOLOv3 and YOLOv4: Multiple Object Detection for Surveillance Applications," *Proceedings of the Third International Conference on Smart Systems and Inventive Technology,* 2020.

[13] A. Bochkovskiy, C.-Y. Wang and H.-Y. Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv:2004.10934v1 [cs.CV],* 2020.

[14] P. K. Mishra and G. P. Saroha, "A Study on Video Surveillance System for Object Detection and Tracking," *IEEE,* 2016.

[15] A. Bari, S. Waseem, S. and S., "Social Distancing Through Image Processing, Video Analysis, and CNN," in *International Conference on*

Computational Intelligence and Emerging Power System, 2022.

[16] A. H. Ahamad, N. Zaini and M. F. A. Latip, "Person Detection for Social Distancing and Safety Violation Alert based on Segmented ROI," *IEEE International Conference on Control System, Computing and Engineering (ICCSCE2020),* 2020.

[17] D. k. R. S. M. P. V. R. D. P. R. Harish Adusumalli, "If the input is a video stream, the picture or a frame of the video is initially delivered to the default face detection module for detection of human faces. This is accomplished by first enlarging the picture or video frame, then identifying the blob ins," *Proceeding of the third international conference on intelligent communication technologies and virtual mobile networks,* 2021.

[18] H. Adusumalli, D. Kalyani and R. K. Sri, "Face Mask Detection Using OpenCV," *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021).,* 2021.