



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic

Mohammad Shorfuzzaman^a, M. Shamim Hossain^{b,*}, Mohammed F. Alhamid^b

^a Department of Computer Science, College of Computers and Information Technology (CCIT), Taif University, Taif, Saudi Arabia

^b Department of Software Engineering, College of Computer and Information Sciences, King Saud University, P.O. Box: 51178, Riyadh 11543, Saudi Arabia

ARTICLE INFO

Keywords:

Sustainable cities
COVID-19 pandemic
video surveillance
social distancing
deep learning
object detection

ABSTRACT

Sustainable smart city initiatives around the world have recently had great impact on the lives of citizens and brought significant changes to society. More precisely, data-driven smart applications that efficiently manage sparse resources are offering a futuristic vision of smart, efficient, and secure city operations. However, the ongoing COVID-19 pandemic has revealed the limitations of existing smart city deployment; hence, the development of systems and architectures capable of providing fast and effective mechanisms to limit further spread of the virus has become paramount. An active surveillance system capable of monitoring and enforcing social distancing between people can effectively slow the spread of this deadly virus. In this paper, we propose a data-driven deep learning-based framework for the sustainable development of a smart city, offering a timely response to combat the COVID-19 pandemic through mass video surveillance. To implementing social distancing monitoring, we used three deep learning-based real-time object detection models for the detection of people in videos captured with a monocular camera. We validated the performance of our system using a real-world video surveillance dataset for effective deployment.

1. Introduction

Due to the coronavirus disease 2019 (COVID-19), the world is undergoing a situation unprecedented in recent human history, with massive economic losses and a global health crisis. The virus initially identified in December 2019 in the city of Wuhan, China has rapidly spread throughout the world, resulting in the ongoing pandemic. Since the initial outbreak, the disease has affected over two hundred countries and territories across the globe, with more than 20 million cases reported (COVID-19, 2020). The outbreak was declared a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO) (WHO, 2020) on January 30, 2020. The virus is very contagious and is primarily transmitted between people through close contact. A variety of common symptoms are found in those infected, such as cough, fever, shortness of breath, fatigue, loss of smell, and pneumonia. The complications of the disease include pneumonia, acute respiratory distress syndrome, and other infections. Precise and timely diagnosis is being hampered due to the lack of treatment, scarcity of resources, and harsh conditions of the laboratory environment. This has increased the challenge to curb the spread of the virus.

Furthermore, the absence of an approved therapy to cure COVID-19

infections has motivated the pressing need for prevention and mitigation solutions to reduce the spread of the virus. Social distancing protocols, including country-wide lockdowns, travel bans, and limiting access to essential businesses, are gradually curbing the spread. In fact, social distancing has already proven to be an effective non-pharmaceutical measure for stopping the transmission of this infectious disease (Ferguson et al., 2006). Social distancing refers to an approach to minimizing disease spread by maintaining a safe physical distance between people, avoiding crowds, and reducing physical contact. According to WHO norms (Hensley, 2020), proper social distancing requires people to maintain a distance of at least 6 ft from other individuals. Because it is highly likely that an infected individual may transmit the virus to a healthy person, social distancing can significantly reduce the number of fatalities caused by the virus, as well as reduce economic loss. Fig. 1 illustrates the impact of social distancing on the daily number of cases (Irfan, 2020). It can be observed from Fig. 1 (a) that social distancing can significantly reduce the peak number of cases of infection, and essentially delay the occurrence of the peak if it is implemented at an early stage of the pandemic. This would reduce the burden on health care facilities and allow more time for adopting countermeasures. Also, as shown in Fig. 1 (b), social distancing can reduce the total number of

* Corresponding author at: Department of Software Engineering, College of Computer and Information Sciences, King Saud University, P.O. Box: 51178, Riyadh 11543, Saudi Arabia

<https://doi.org/10.1016/j.scs.2020.102582>

Received 10 August 2020; Received in revised form 13 October 2020; Accepted 27 October 2020

Available online 5 November 2020

2210-6707/© 2020 Elsevier Ltd. All rights reserved.

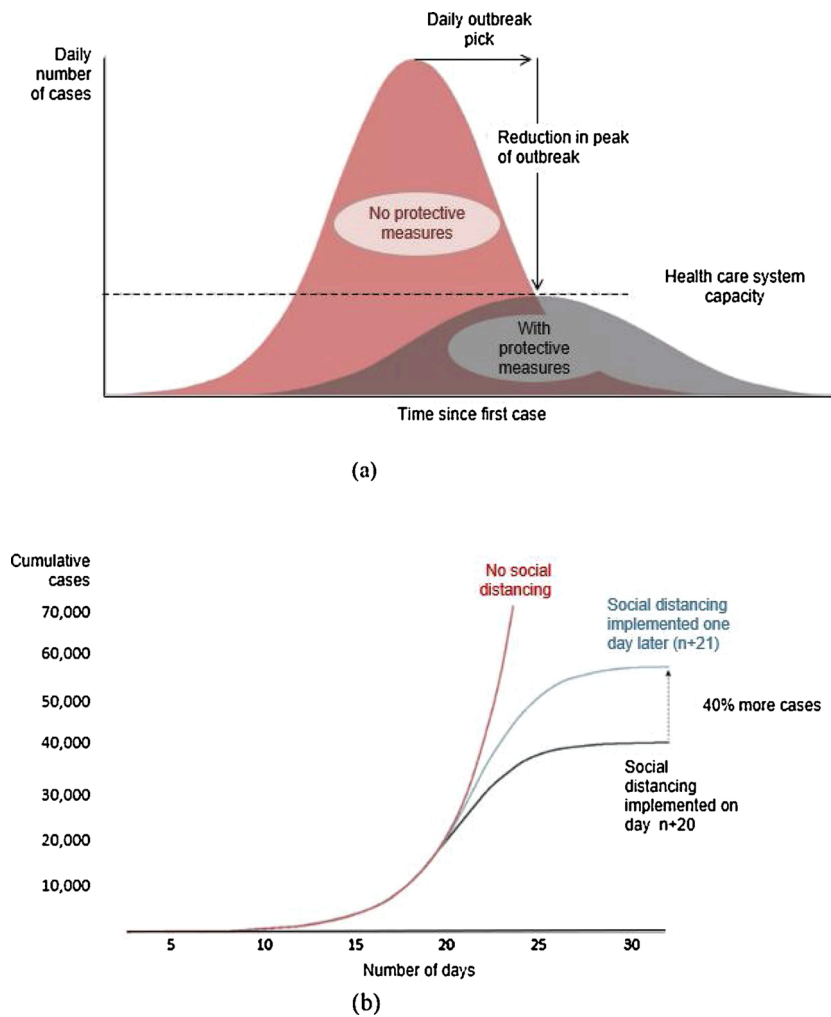


Fig. 1. Impact of social distancing on COVID-19 outbreak [20].

cases, and the sooner the measure is taken, the higher the positive impact will be.

Lately, several countries throughout the world including the Netherlands (Amsterdam, 2020), USA (Smart America, 2020), and South Korea (Silva, Khan, & Han, 2018) are taking the initiative to deploy “sustainable smart cities” (Bibri & Krogstie, 2017). For example, the latest Smart City 3.0 (Amsterdam, 2020) initiative by the City of Amsterdam encourages the effective participation of citizens, government, and private organizations in building smart city solutions. The plan includes the development of infrastructures and technologies in the areas of smart energy and water systems, the Intelligent Transport System (ITS), and so on. However, as part of the effective preparation for the current and future pandemics, it is expected that the sustainable development of smart cities will provide situational intelligence and an automated targeted response to ensure the safety of global public health and to minimize massive economic losses. In this context, smart cities will host data-driven services along with other IoT devices, such as IP surveillance and thermal cameras, sensors, and actuators, to deliver community-wide social distancing estimates and the early detection of potential pandemics. Fig. 2 illustrates a sustainable smart city scenario where social distancing is monitored in real time to offer a variety of services, such as detecting and monitoring the distance between any two individuals, detecting crowds and gatherings in public areas, monitoring physical contacts between people such as handshaking and hugging, detecting and monitoring individuals with disease symptoms such as cough and high body temperature, and monitoring any violation of quarantine by infected people. In this paper, we propose a data-driven

deep learning framework for the development of a sustainable smart city, offering a timely response to combat the COVID-19 pandemic through mass video surveillance. Upon the detection of a violation, an audio-visual, non-intrusive alert is generated to warn the crowd without revealing the identities of the individuals who have violated the social distancing measure. In particular, we make the following contributions: (a) A deep learning-based framework is presented for monitoring social distancing in the context of sustainable smart cities in an effort to curb the spread of COVID-19 or similar infectious diseases; (b) The proposed system leverages state-of-the-art, deep learning-based real-time object detection models for the detection of people in videos, captured with a monocular camera, to implement social distancing monitoring use cases; (c) A perspective transformation is presented, where the captured video is transformed from a perspective view to a bird’s eye (top-down) view to identify the region of interest (ROI) in which social distancing will be monitored; (d) A detailed performance evaluation is provided to show the effectiveness of the proposed system on a video surveillance dataset.

The rest of the paper is organized as follows. The background and related work is presented in Section 2. Sections 3 and 4 present the proposed system, dataset, and experiments with the performance results. Finally, Section 5 concludes the paper with suggestions for future work.

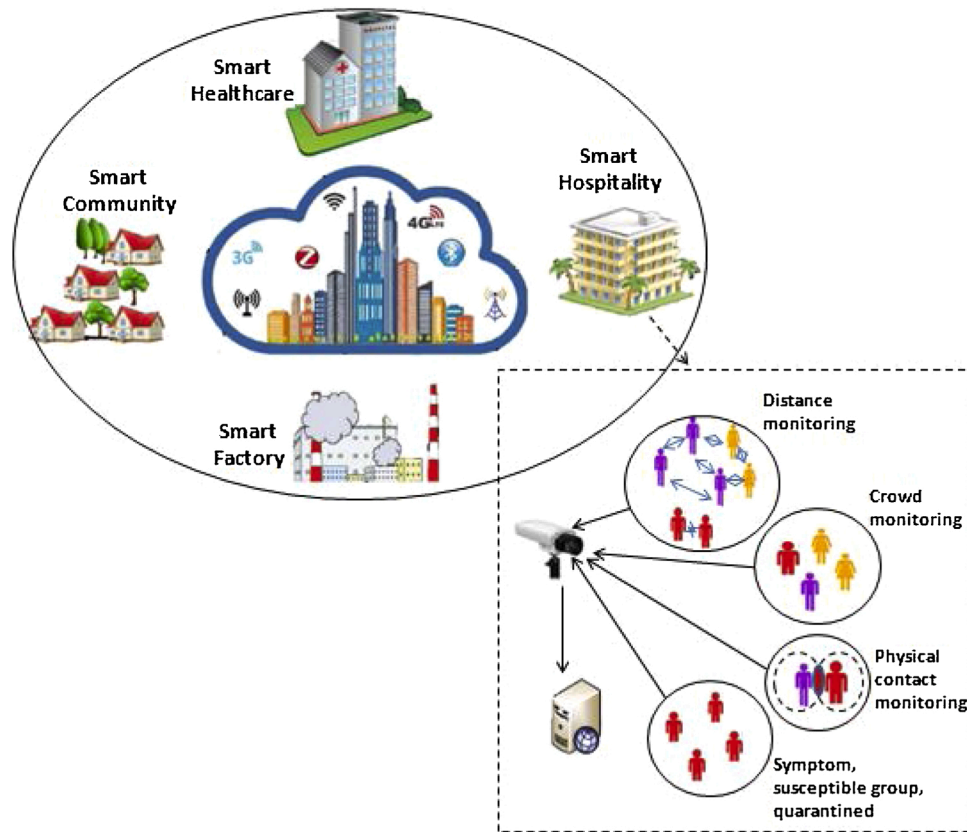


Fig. 2. Real-time monitoring of social distancing in a sustainable smart city scenario.

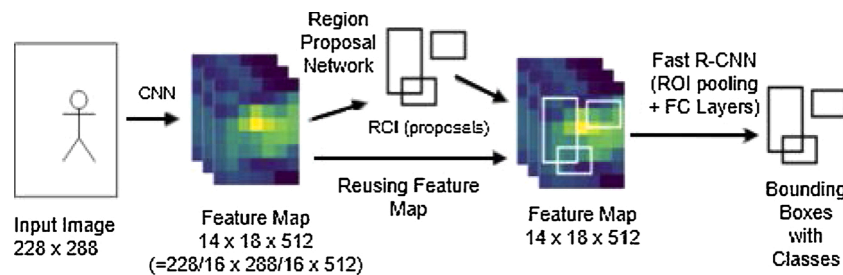


Fig. 3. Architectural representation of Faster R-CNN.

2. Background and Related Studies

2.1. State-of-the-Art Object Detection Models

Object detection is one of the most challenging problems in the computer vision domain, and lately there has been substantial improvement in this field with the advancements in deep learning (Yang et al., 2016). In this study, we use three state-of-the-art object detection architectures that are pre-trained and optimized on large image datasets, such as PASCAL-VOC (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010) and MS-COCO (Lin et al., 2014), to detect pedestrians for monitoring social distancing in mass video surveillance footage through vision-based social media event analysis (Qian, Zhang, Xu, & Hossain, 2015; Yang, Zhang, Xu, & Hossain, 2015; Alhamid et al., 2015). We present a two-stage detector called Faster R-CNN (Faster Region with Convolutional Neural Networks) (Ren, He, Girshick, & Sun, 2015) and two one-stage detectors called SSD (single shot multistage detector) (Liu et al., 2015) and YOLO (you only look once) (Redmon & Farhadi, 2017).

2.1.1. Faster R-CNN

Faster R-CNN (Ren et al., 2015) was built incrementally from two of its predecessor architectures, called R-CNN (Girshick, Donahue, Darrell, & Malik, 2014) and Fast R-CNN (Girshick, 2015), where ROIs are generated using a technique called selective search (SS). Because SS does not involve any deep learning techniques, the authors of Faster R-CNN proposed the region proposal network (RPN), which uses CNN models such as ResNet 101 (He, Zhang, Ren, & Sun, 2016), VGG-16 (Simonyan & Zisserman, 2015), and Inception v2 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) to generate region proposals. This increases the speed of the Faster R-CNN compared with Fast R-CNN at least tenfold. Fig. 3 shows a schematic diagram of Faster R-CNN architecture, where the RPN accepts an image as input and outputs an ROI. Each ROI consists of a bounding box and an objectness probability. To generate those numbers, a CNN is used to extract a feature volume. After post-processing, the final output is a list of ROIs. In the second stage, Faster R-CNN performs classification in which it accepts two inputs, namely the list of ROIs from the previous step (the RPN) and a feature volume computed from the input image, and outputs the final bounding boxes.

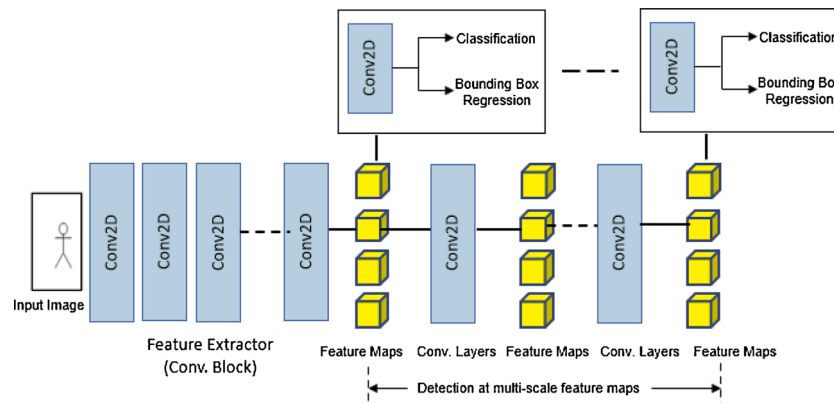


Fig. 4. Architectural representation of SSD.

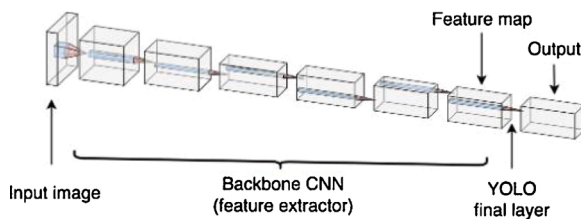


Fig. 5. Architecture summary of YOLO.

2.1.2. SSD

Researchers from Google introduced the SSD architecture (Liu et al., 2015) which performs object detection in real time by combining region proposal and feature extraction in a single deep neural network. SSD overcomes the limitation of Faster R-CNN, which shows better accuracy but suffers from a slow frames per second (FPS) rate. SSD produces bounding boxes in conjunction with scores based on the objects detected in those boxes, followed by a post-processing non-maxima suppression (NMS) step to generate the final detections. Consequently, the detection process consists of two steps, namely feature map extraction and object detection through convolutional filtering built from three separate components. The first part represents the base pre-trained network (such as MobileNet) (Howard et al., 2017), which is used for feature extraction. The second part consists of a series of convolutional filters representing multi-scale feature layers. Finally, an NMS unit represents the last layer, where unwanted overlapping bounding boxes are removed to produce only one box per object. A schematic architecture of SSD is shown in Fig. 4.

2.1.3. YOLO

Another single stage object detector is YOLO (shown in Fig. 5) which is often considered a competitor of SSD. We used YOLOv3 in this study. It is one of the fastest object detection algorithms available in the literature, and can run at more than 170 FPS on a modern GPU. However, it is outperformed by Faster R-CNN in terms of accuracy. Moreover, due to the way it detects objects, YOLO struggles with smaller objects. Nevertheless, the architecture is constantly evolving from its earlier variants (Redmon & Farhadi, 2017), and its challenges are being worked on. The core idea of YOLO is that it reframes object detection as a single regression problem. The model is split into two parts, namely inference and training. Inference refers to the process of taking an input image and computing results, while training represents the process of learning the weights of the model. Like most other image detection models, YOLO is based on a backbone model that extracts meaningful features from the image to be used in the final layers. While any architecture can be chosen as a feature extractor, the YOLO study employs a custom architecture called Darknet-53. The performance of the final

model depends heavily on the choice of feature extractor architecture.

2.2. Monitoring Social Distancing

Since the onset of the COVID-19 pandemic, many countries around the world have taken the initiative to develop solutions for combatting the outbreak based on emerging technology. Many law enforcement departments are making use of drones and video surveillance cameras to detect and monitor crowded areas and adopt disciplinary actions that alert the crowd (Robakowska, Tyranska-Fobke, Nowak, & Slezak, 2017). A recent study (Nguyen, Saputra, & Van Huynh, 2020) investigated how social distancing can be enforced through various scenarios, and by using technologies such as AI and IoT. The authors used the basic concept of social distancing, and various models that used existing technologies to control the spread of the virus. Agarwal et al. (Agarwal et al., 2020) discussed state-of-the-art disruptive technologies to fight the COVID-19 pandemic. They introduced the notion of disruptive technologies and classified their scope in terms of human-centric or smart-space categories. Furthermore, the authors provided a SWOT analysis of the identified techniques. Khandelwal, Khandelwal, and Agarwal. (2020) proposed a computer vision-based system to monitor the activities of a workforce to ensure their safety using CCTV feeds. As part of the system, they built tools to effectively monitor social distancing and to detect face masks.

Another recent work presented by Pun, Sonbhadra, and Agarwal (2020) proposed a social distancing monitoring approach using YOLOv3 and Deep SORT to detect pedestrians and calculate a social distancing violation index. The study was limited by the lack of statistical analysis and direction for deployment. Cristani, Del Bue, Murino, Setti, and Vinciarelli (2020) also proposed a special social distancing monitoring approach in which they formulated the monitoring problem as visual social distancing (VSD) problem. They discussed the impact of the subjects' social context on the computation of distances, and they raised privacy concerns. Hossain, Muhammad, and Guizani (2020) presented a health care framework based on a 5 G network to develop a mass video surveillance system for monitoring body temperature, face masks, and social distancing. Sun and Zha (2020) introduced and developed two critical indices called social distance probability and ventilation effectiveness for the prediction of COVID-19 infection probability. Using these indices, the authors demonstrated the impact of social distancing and ventilation on the risk of respiratory illness infection. Rahman, Zaman, Asyhari, and Al-Turjman (2020) presented a data-driven approach to building a dynamic clustering framework to alleviate the adverse economic impact of COVID-19. They developed a clustering algorithm to simulate various scenarios, and thus to identify the strengths and weaknesses of the algorithm. Kolhar, Al-Turjman, Alameen, and Abualhaj (2020) proposed a decentralized IoT-based system to regulate population flow during the COVID-19 lockdown.

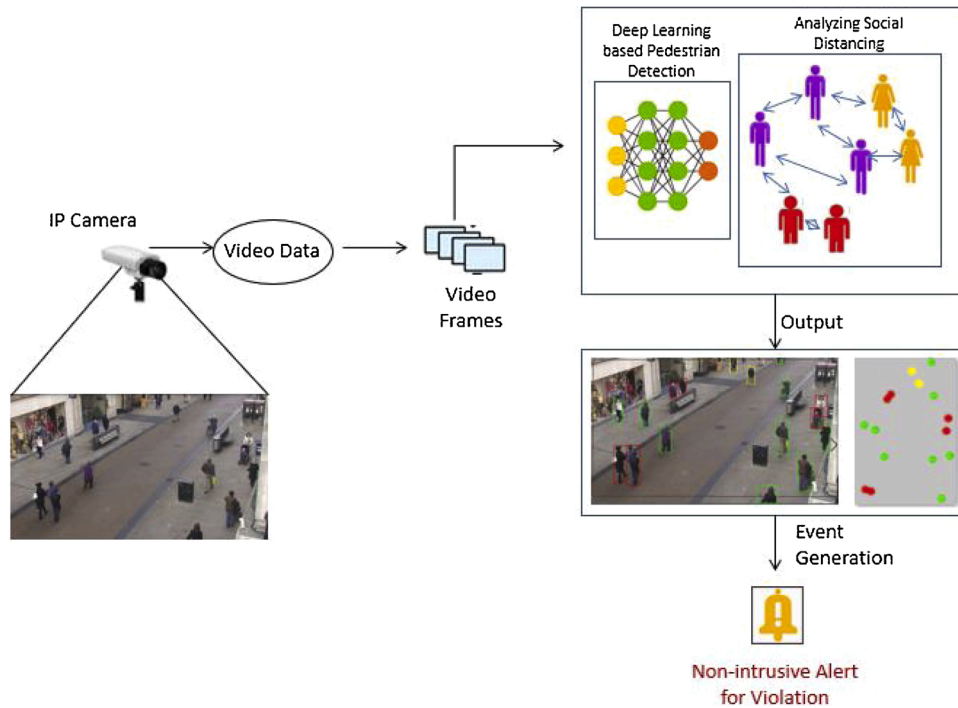


Fig. 6. Illustration of the proposed system. Real-time video data from an IP surveillance camera is directly fed into the system for social distancing monitoring. An audio-visual non-intrusive dismissible alert is generated for any violation.

The system used state-of-the-art face detection and recognition methods in both centralized and distributed architectures. The experimental results demonstrated the superiority of decentralized face detection in cloud and edge computing. Yassine and Hossain (2020), presented an auction-based model for allocating network resources to battle the COVID-19 crisis. In another effort, Al-Turjman and Deebak (2020) presented a privacy-aware framework to combat the ongoing pandemic in an energy-efficient manner using the Internet of Medical Things (IoMT). A social distancing monitoring framework called DeepSOCIAL was

proposed by Rezaei and Azarmi (2020) that used a generic deep neural network-based model to automatically detect people in a crowded area through CCTV security cameras. The authors also developed an online risk assessment scheme using spatio-temporal data to identify zones with high risk of potential spread and infection of the virus. Sathya-moorthy, Patel, Savle, Paul, and Manocha (2020) proposed a social distancing monitoring scheme based on a mobile robot with commodity sensors that could navigate through a crowd without collision to estimate the distance between all detected people. The robot was also

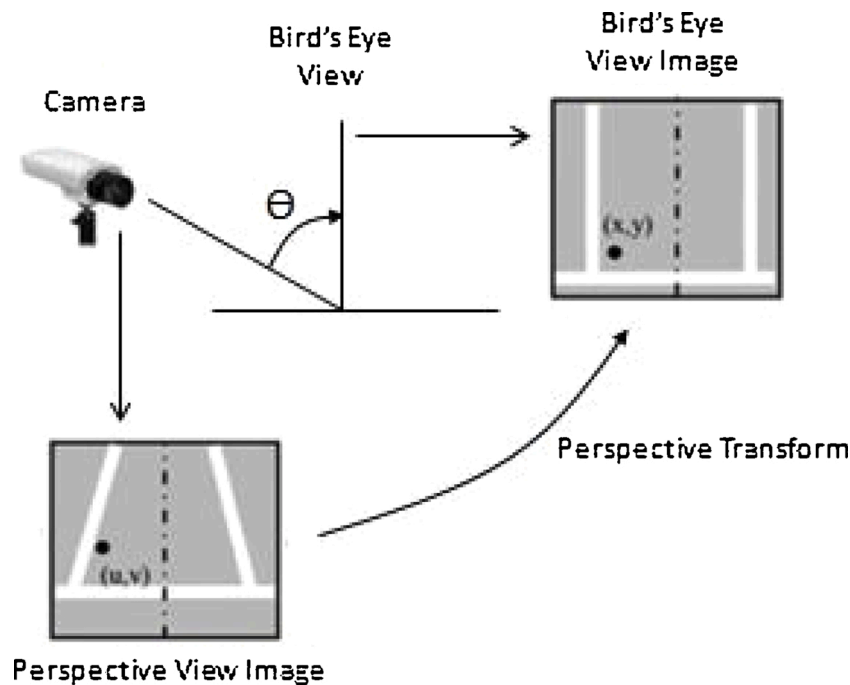


Fig. 7. Perspective transformation representation.

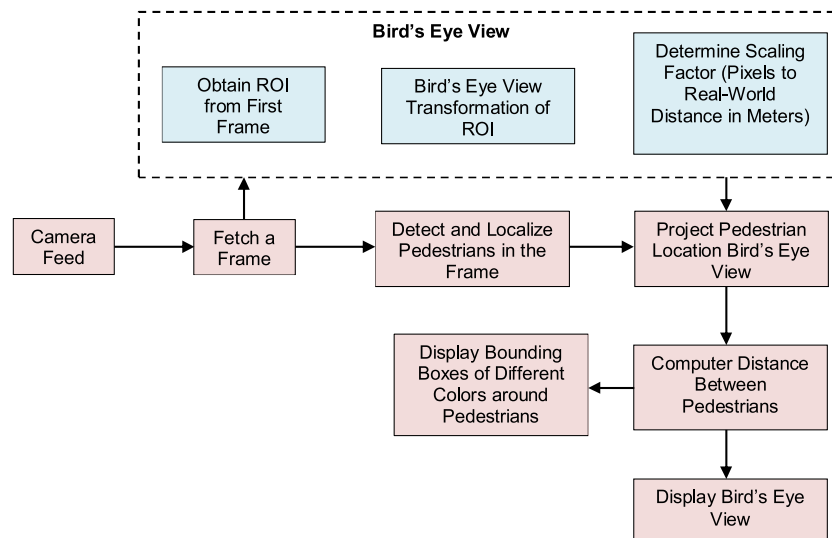


Fig. 8. Algorithmic flow of the proposed system. OpenCV's perspective transform routine is used for bird's eye view transformation.

equipped with thermal cameras to remotely transmit thermal images to security personnel who monitored individuals with a higher-than-normal temperature. Fan et al. (2020) presented a similar approach to social distancing monitoring with an autonomous surveillance quadruped robot that could promote social distancing in complex urban environments.

The existing systems in the literature that leverage various measurements for social distancing monitoring are interesting, however, recording and storing surveillance data and generating intrusive alerts may not be acceptable for many individuals. Hence, the current implementation of the proposed system detects pedestrians in an ROI using a fixed monocular camera and estimates the distance between pedestrians in real time without recording data. Our system generates an audio-visual, non-intrusive dismissal alert to caution the crowd when it detects any social distancing violation. Moreover, a perspective transformation is presented, where the captured video is transformed from a perspective view to a bird's eye (top-down) view to determine the ROI in which social distancing will be monitored.

3. Proposed Framework

The recent advancement in deep learning technology has brought significant improvement to the development of techniques for a broad range of challenges and tasks involved in medical diagnosis (Punn & Agarwal, 2020), epilepsy seizure detection (Hossain, Amin, Muhammad, & Al Sulaiman, 2019), speech recognition (Amodei, Ananthanarayanan, Anubhai, & Bai, 2016), machine translation (Vaswani, Bengio, Brevdo, Chollet, & Gomez, 2018), and so on. The majority of these tasks are focused on classification, segmentation, detection, recognition, and the tracking of objects (Brunetti, Buongiorno, Trotta, & Bevilacqua, 2018; Punn & Agarwal, 2019). To this end, the state-of-the-art CNN-based architectures pre-trained and optimized on large image datasets such as PASCAL-VOC (Everingham et al., 2010) and MS-COCO (Lin et al., 2014) have shown substantial performance improvement for object detection. Motivated by this, we present in this study a deep learning-based video surveillance framework using state-of-the-art object detection and tracking models to monitor physical distancing in crowded areas in an attempt to combat the COVID-19 pandemic. For the sake of simplicity, the current implementation of the proposed system detects pedestrians in an ROI using a fixed monocular camera and estimates the distance between pedestrians in real time without recording data. Recording and storing surveillance data and generating intrusive alerts may not be acceptable by many individuals. Our system generates an audio-visual

non-intrusive dismissal alert to signal the crowd upon detecting any social distancing violation. A general overview of our system is presented in Fig. 6, and a detailed description starts below.

3.1. Perspective Transformation

The incoming video may be fed to the system from any perspective view, and hence we first needed to transform the video from a perspective view to a bird's eye (top-down) view. To achieve this, we selected four points in the perspective view that formed the ROI where social distancing would be monitored. Subsequently, we align these four points to the four corners of a rectangle in the bird's eye view. Fig. 7 illustrates an intuitive representation of perspective transformation reproduced from the study by Luo, Koh, Min, Wang, and Chong (2010). After the transformation, the concerned points constitute parallel lines if they are observed from the top (hence the bird's eye view). This bird's eye view is characterized by a uniform distribution of points in both horizontal and vertical directions, even though the scale is different in each direction. We also measured the scaling factor of the bird's eye view during this calibration process, by which we determined how many pixels should correspond to 6 ft in real-world coordinates. Thus, we can obtain a transformation that can be applied to the entire image in perspective view.

3.2. Pedestrian Detection

In the second step, we detect pedestrians in the transformed image view with the selected object detection models (Faster R-CNN, YOLO, SSD) trained on real-world datasets. Subsequently, a bounding box with four corners is drawn for each detected pedestrian. Multiple detections of an individual pedestrian result in multiple overlapping bounding boxes. We use non-max suppression (NMS) to remove unwanted bounding boxes to ensure that our detector detects a pedestrian only once.

3.3. Inter-Pedestrian Distance Calculation

The last step is to calculate the distance between each pair of pedestrians to detect any potential violation of the social distancing norm. To do this, we make use of the bounding box for each pedestrian in the image. To localize the detected pedestrian in the image, we take the bottom center point of the bounding box and apply a perspective transformation on it, resulting in a bird's eye view of the position of the

detected pedestrian. After calculating the distance between every pair of pedestrians in the bird's eye view, we identify the pedestrians whose distance is below the minimum acceptable threshold and highlight them with red bounding boxes, and at the same time generate a non-intrusive audio-visual alert to warn the crowd. Based on the calculated distance, other pedestrians are marked as safe or at low risk with green and yellow, respectively. The complete algorithmic flow of the detection process is shown in Fig. 8.

4. Experiments and Results Analysis

To demonstrate the effectiveness of our video surveillance framework while monitoring social distancing in crowded areas, we extensively evaluated the proposed framework using all three object detection models—Faster R-CNN, YOLO, and SSD—with the publicly available Oxford Town Center dataset (Benfold & Reid, 2011). This is a video dataset that was released by Oxford University as part of the visual surveillance project. It contains video data from a semi-crowded town center (an urban street) with a resolution of 1920×1080 sampled at 25 FPS. The video was downsampled to a standardized resolution of 1280×720 before it was fed to the object detection models. The dataset also contains the ground truth bounding boxes for the pedestrians in all the frames in the entire video. We evaluated the object detection models for person detection in the test video using the predicted bounding boxes and the coordinates from the ground truth boxes.

The implementation started with obtaining the perspective transformation (top-down view) of the video. We used a mouse click event to select the ROI, where we chose four points to designate the area in the first frame to monitor the social distancing. This is a one-time process that was repeated for all the frames in the video. Next, three points were chosen to define a 6 ft (approximately 180 cm) distance in both the vertical and horizontal directions, forming lines parallel to the ROI. From these three points, a scaling factor was calculated for use in the top-down (bird's eye) view in both directions to determine how many pixels corresponded to 6 ft in real-world coordinates. In the second step, we applied object detection models to detect pedestrians and draw a bounding box around each of them. As mentioned, we applied NMS and other rule-based heuristics as part of the minimal post-processing on the output bounding boxes to reduce the possibility of over-fitting. After the pedestrians were located, their positions were transformed into real-world coordinates through bird's eye view transformation. The pre-trained object detection models optimized on MS COCO (Lin et al., 2014) and PASCAL VOC (Everingham et al., 2010) datasets were implemented using PyTorch and TensorFlow. More particularly, the Detectron2 API from the PyTorch and TensorFlow object detection API was used. We conducted experiments in the Google Colab notebook environment, which provides free GPU access. It currently offers an NVIDIA Tesla P100 GPU with 16 GB RAM, and is equipped with pre-installed Python 3.x packages, PyTorch, and the Keras API with a TensorFlow backend.

In the third step, we conducted social distancing monitoring by calculating the distance between each pair of pedestrians by measuring from the bottom center point of each pedestrian's bounding box. The statistics related to the total number of violations and the level of risk for individuals were recorded over time. In subsequent sections, we present the metrics for evaluation and the experimental results with discussion.

4.1. Evaluation Metrics

For various annotated datasets such as PASCAL VOC (Everingham et al., 2010) and MS COCO (Lin et al., 2014), and their relevant object detection challenges, the most widely used performance metric for estimating detection accuracy was the average precision (AP). In this study, we use similar metrics to demonstrate the performance of our social distancing framework. In particular, the object detection metrics provide an estimate of how well our model performs on a person

$$J(\text{bbox}_{gt}, \text{bbox}_p) = \text{IoU} = \frac{\text{area}(\text{bbox}_{gt} \cap \text{bbox}_p)}{\text{area}(\text{bbox}_{gt} \cup \text{bbox}_p)} \quad (1)$$

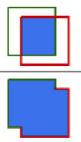
$$\text{IoU} = \frac{\text{area of overlap}}{\text{area of union}} = \frac{\text{img}}{\text{img}}$$


Fig. 9. Illustrating intersection over union (IoU).

detection task in mass surveillance areas. In this context, it is important to distinguish between correct and incorrect detections. A common way to do this is to use the intersection over union (IoU) metric. IoU, also referred to as the Jaccard Index, is used to measure the similarity between two datasets (Jaccard, 1901). In the context of object detection, it provides a measure of the similarity between the ground truth bounding box and the predicted bounding box as a measurement for the quality of the prediction. The value of IoU varies from 0 to 1. The closer the bounding boxes, the higher the value of IoU. Specifically, the IoU estimates the overlap of ground truth (bbox_{gt}) and predicted (bbox_p) bounding boxes over the area obtained by their union, as illustrated in the following equation and Fig. 9.

$$J(\text{bbox}_{gt}, \text{bbox}_p) = \text{IoU} = \frac{\text{area}(\text{bbox}_{gt} \cap \text{bbox}_p)}{\text{area}(\text{bbox}_{gt} \cup \text{bbox}_p)} \quad (1)$$

Now, after computing the IoU for each detection, we compared it with a given threshold, T_{th} to obtain a classification for the detection. If the value of IoU was above the threshold, the detection was considered as a positive (correct) prediction. On the contrary, if the value of IoU was below the threshold, the detection was considered as a false (incorrect) prediction. More specifically, the predictions were categorized as true positive (TP), false positive (FP), and false negative (FN). Intuitively, there are two cases that are deemed as FPs. In one, the object is present but the IoU is less than the threshold, and in the other case, the object is not present, but the model detects one. FN refers to the case where the object is present, but the model fails to detect it.

Based on these various prediction types, precision and recall values were calculated and served as the basis for creating precision \times recall curves and computing mean AP (mAP). Precision refers to the model's ability to detect relevant objects and was calculated as the percentage of correct detections over all positive detections. Recall refers to the model's sensitivity and was calculated as the percentage of correct positive predictions over all ground truth objects. The precision \times recall curve summarizes both precision and recall as a trade-off for various confidence values linked to the bounding boxes produced by the detection model. In practice, the curve appears to be very noisy due to the trade-off between precision and recall, and hence it is difficult to estimate the model performance by computing the area under the curve (AUC). This is managed by smoothing out the curve before AUC estimation by means of a numerical value called AP. There are two different techniques, called 11-point and all-point interpolation, used to achieve this. In fact, the computation method for AP was changed by the PASCAL VOC challenge (Everingham et al., 2010) from 2010 onward. At present, all data points are used for interpolation, rather than interpolating at only 11 points that are equally spaced. However, we adopted both interpolation techniques for the sake of completeness.

4.1.1. Point Interpolated AP

This approach summarizes the precision and recall curve by taking an average of the maximum precision values across a set of 11 equally spaced recall values in the range of 0 to 1. More precisely, we interpolated the precision score for a certain recall value, r , by taking the maximum precision where the corresponding recall value, \tilde{r} was greater than r . This can be formulated as follows:

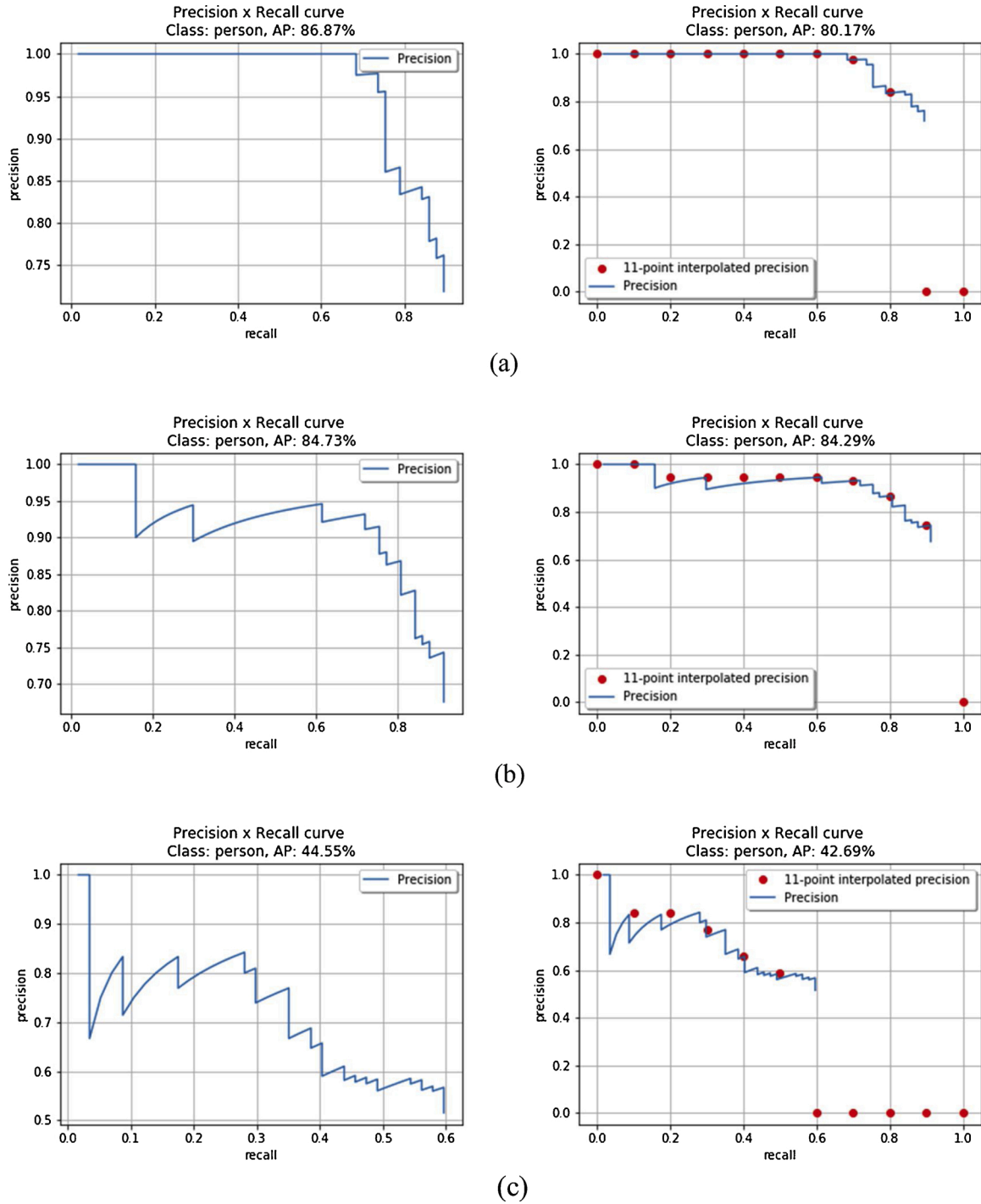


Fig. 10. Precision and recall curves based on all-point and 11-point interpolation for pedestrian detection using various object detection models at $IoU=0.5$: (a) Faster R-CNN, (b) YOLO, and (c) SSD.

$$AP_{11} = \frac{1}{11} \sum_{r \in \{0, 0.1, 0.2, \dots, 1\}} P_{interp}(r) \quad (2)$$

where the interpolated precision is denoted as:

$$P_{interp}(r) = \max_{\tilde{r} > r} P(\tilde{r}) \quad (3)$$

4.1.2. All-Point Interpolated AP

In this case, we compute AP by interpolating the precision score at all recall values instead of using only 11 recall levels. This can be translated mathematically as follows:

$$AP_{all} = \sum_n (r_n - r_{n+1}) P_{interp}(r_n) \quad (4)$$

where the interpolated precision is denoted as:

$$P_{interp}(r_n) = \max_{\tilde{r} > r_n} P(\tilde{r}) \quad (5)$$

Now that we have defined the AP, the mean average precision (mAP) simply refers to the mean of the APs across all classes (N) in the dataset, and is calculated as follows:

Table 1
Comparison of performance for various object detection models.

Model	mAP	meanIoU	FPS	IoU Threshold
Faster R-CNN	0.868	0.907	4	0.7
YOLO	0.847	0.906	9	0.7
SSD	0.445	0.918	25	0.2

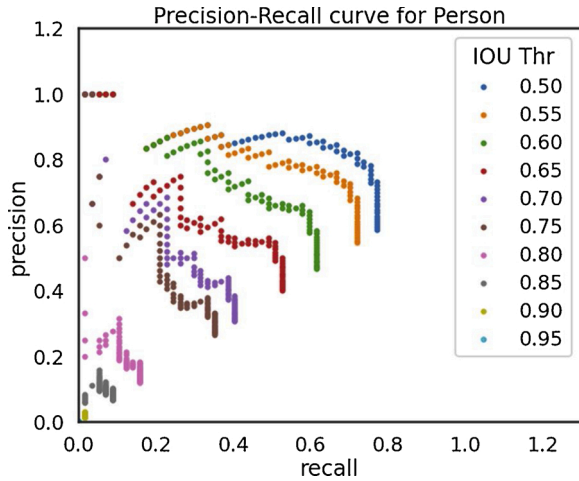


Fig. 11. (a) Precision and recall curves for pedestrian detection using the YOLO object detection model over different IoU thresholds 0.50:0.05:0.95 [AP@[0.5,0.95]].

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \tag{6}$$

4.2. Person Detection Evaluation

We used three different CNN-based object detection models, namely Faster R-CNN, YOLOv3, and SSD, for experiments with social distancing monitoring. Fig. 10 illustrates the precision-recall curves and AP for all three detectors using both all-point and 11-point interpolation techniques computed at a single IoU threshold of 0.5 (standardized IoU value for the PASCAL VOC ((Everingham et al., 2010) evaluation). As expected, Faster R-CNN showed the best AP results using both interpolation techniques. YOLOv3 achieved AP values that were remarkably close to Fast R-CNN, while SSD-MobileNet appears to be the least accurate model from the list for the test townhouse dataset. In fact, SSD-MobileNet performed reasonably well when used for detecting large objects close by, but showed poor performance with this dataset, which contains many pedestrians, each occupying a small view space. We reduced the IoU threshold significantly (as shown in Table 1) to obtain reasonable detection results for the test video footage. Actually, the IoU threshold is a tunable parameter that can be adjusted to control the values of precision and recall, which have a direct impact on a model’s detection performance. From the precision-recall curve, we calculated AP by computing AUC, which represents the aggregated areas under the curve, where each distinct area is specified by a decline in precision value at a specific level of recall. Furthermore, Fig. 11 shows the AP scores for the YOLOv3 object detection model over various IoU threshold values starting at 0.5. A declining trend in AP values with higher IoU thresholds is generally

observed. Finally, Table 1 summarizes the performance of various object detection models for pedestrian detection with the test townhouse surveillance video footage. Considering both accuracy and the speed of

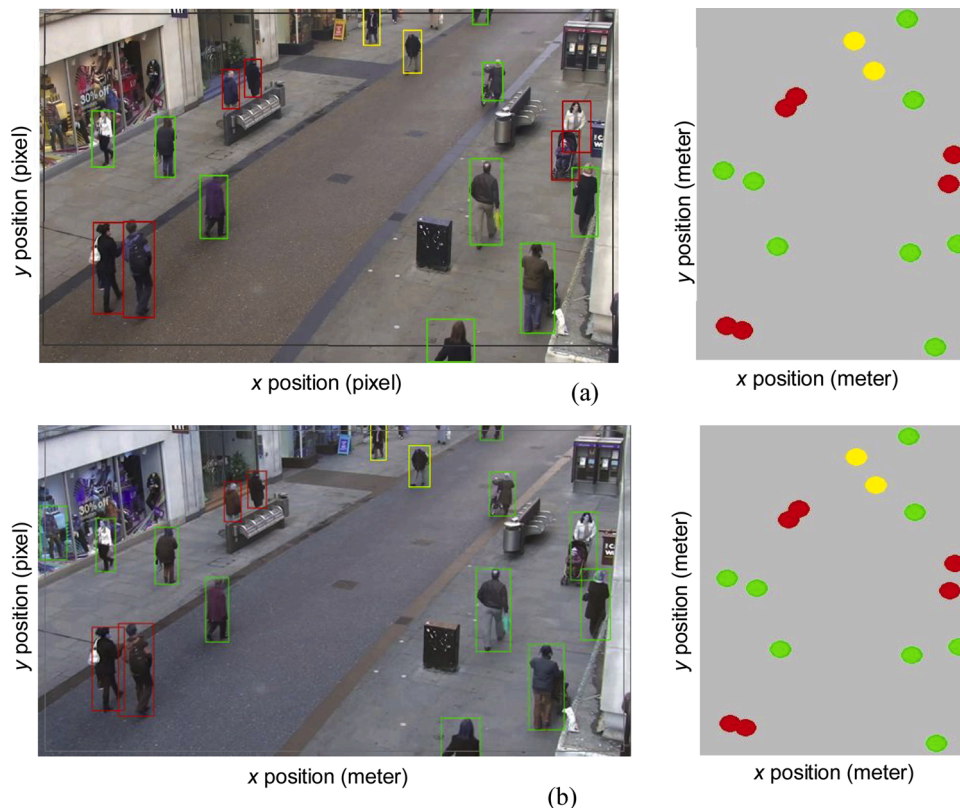


Fig. 12. Illustrating pedestrian detection using (a) YOLO (b) Faster R-CNN and social distancing monitoring (left: surveillance footage with bounding boxes, right: top-down transformation (bird’s eye view) showing violations in red).

detection (in FPS), YOLO appears to be the best model for obtaining a fair trade-off between these two performances metrics. Hence, in our framework, we opted to use the YOLO object detection model for social distancing monitoring on surveillance video. In addition, relatively higher values of mean IoU for all detection models indicate that pedestrians detected in the video footage were perfectly localized by the bounding boxes.

4.3. Qualitative Evaluation

Fig. 12 illustrates pedestrian detection screenshots using YOLOv3 and Faster R-CNN models, and the corresponding social distancing in real-world transformed bird's eye views. The processed video frames (in pixel coordinates) with detected pedestrians confined to bounding boxes of different colors are shown on the left. The corresponding top-down transformations in real-world coordinates (in meter coordinates) are shown on the right, where various levels of risks are identified with distinct colors. Social distancing violations, low risk, and safe distances are labeled with red, yellow, and green color, respectively. In general, the Faster R-CNN models appear to be overly sensitive and detected a plastic human display as a pedestrian, as shown in Fig. 4 (b).

5. Conclusions

For sustainable development, maintaining safety and encouraging well-being at all ages is important. The current pandemic has devastated the sustainable development of society. This study is a step toward a better understanding of the dynamics of the COVID-19 pandemic, and proposes a deep learning-based solution for the efficient monitoring of social distancing through mass video surveillance in a sustainable smart city scenario for the well-being of the citizens. The proposed system aims to achieve this through state-of-the-art deep learning-based object detection models to detect and track individuals in real-time with the help of bounding boxes. Upon the detection of a violation, an audio-visual non-intrusive alert is generated to warn the crowd without revealing the identities of the individuals who have violated the social distancing measure. An extensive performance evaluation was done using Faster R-CNN, SSD, and YOLO object detection models with a public video surveillance dataset, in which YOLO proved to be the best performing model with balanced mAP score and speed (FPS).

The absence of an effective vaccine and the lack of immunity against COVID-19 have made social distancing a largely feasible and widely adopted approach to controlling the ongoing pandemic. Maintaining social distancing has also been recommended by leading health organizations, such as the WHO and Centers for Disease Control and Prevention (CDC). To this end, our proposed deep learning-based video surveillance framework will play a significant role in combating the spread of COVID-19 in a sustainable smart city context. At this stage, it is imperative to identify some of the potential impact of our approach on the surrounding environments, such as increased anxiety and panic among the individuals who receive the repetitive alerts. In addition, some legitimate concerns regarding individual rights and privacy could be raised, and can be effectively handled by obtaining prior consent from individuals and concealing their identities.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors extend their appreciation to the Deputyship for Research & Innovation, "Ministry of Education" in Saudi Arabia for funding this research work through the project number IFKSURG-228.

References

- Agarwal, S., Punn, N. S., Sonbhadra, S. K., Nagabhusan, P., Pandian, K., & Saxena, P. (2020). *Unleashing the power of disruptive and emerging technologies amid COVID 2019: A detailed review*. arXiv preprint arXiv:2005.11507.
- Alhamid, M. F., et al. (2015). Towards context-sensitive collaborative media recommender system. *Multimedia Tools and Applications*, 74, 11399–11428.
- Al-Turjman, F., & Deebak, D. (2020). Privacy-aware energy-efficient framework using Internet of Medical Things for COVID-19. *IEEE Internet of Things Magazine*. DOI: 10.1109/IOTM.0001.2000123.
- Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. in: *International Conference on Machine Learning*, 173–182.
- Amsterdam (2020). Amsterdam Smart City 3.0. URL: <https://smartcityhub.com/governance-economy/amsterdam-better-than-smart/> (Accessed on:08/10/2020).
- Benfold, B., & Reid, I. (2011). Stable multi-target tracking in real-time surveillance video. in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 3457–3464.
- COVID-19. (2020). Dashboard, CoronaBoard, URL: <https://coronaboard.com/>. Retrieved Aug 10, 2020.
- Bibri, S. E., & Krogstie, J. (2017). Smart sustainable cities of the future: An extensive interdisciplinary literature review. *Sustainable Cities and Society*, 31, 183–212.
- Brunetti, A., Buongiorno, D., Trotta, G. F., & Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300, 17–33.
- Cristani, M., Del Bue, A., Murino, V., Setti, F., & Vinciarelli, A. (2020). *The visual social distancing problem*. arXiv preprint arXiv:2005.04813.
- Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2), 303–338.
- Fan, T., Chen, Z., Zhao, X., Liang, J., Shen, C., Manocha, D., et al. (2020). *Autonomous social distancing in urban environments using a quadruped robot*. arXiv preprint arXiv: 2008.08889v1.
- Ferguson, S. K., et al. (2006). Strategies for mitigating an influenza pandemic. *Nature*, 442(no.7101), 448–452.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.
- Girshick, R. (2015). Fast R-CNN. in *Proceedings of the IEEE international conference on computer vision*, 1440–1448.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hensley, L. (2020). *Social distancing is out, physical distancing is in here is how to do it*. Global News–Canada (27 March 2020).
- Hossain, M. S., Muhammad, G., & Guizani, N. (2020). Explainable AI and Mass Surveillance System-Based Healthcare Framework to Combat COVID-19 Like Pandemics. *IEEE Network*, 34(4), 126–132.
- Howard, A. G., Zhu, M., Chen, B., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*. <http://arxiv.org/abs/1704.04861>.
- Irfan, U. (2020). The math behind why we need social distancing, starting right now. *Vox*. Apr. 22, 2020. Accessed on: Aug. 10, 2020. [Online]. Available: <https://www.vox.com/2020/3/15/21180342/coronavirus-covid-19-us-social-distancing>.
- Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, 547–579.
- Khandelwal, P., Khandelwal, A., & Agarwal, S. (2020). *Using computer vision to enhance safety of workforce in manufacturing in a post COVID world*. arXiv preprint arXiv: 2005.05287.
- Kolhar, M., Al-Turjman, F., Alameen, A., & Abualhaj, M. M. (2020). A three layered decentralized IoT biometric architecture for city lockdown during COVID-19 outbreak. *IEEE Access*, 8, 163608–163617. <https://doi.org/10.1109/ACCESS.2020.3021983>
- Lin, T. Y., et al. (2014). Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision – ECCV 2014*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., et al. (2015). SSD: Single shot multibox detector. in *European Conference on Computer Vision*, 21–37.
- Luo, L., Koh, I., Min, K. Y., Wang, J., & Chong, J. (2010). Low-cost implementation of bird's-eye view system for camera-on-vehicle. in *Proc. of International Conference on Consumer Electronics*, 311–312.
- Nguyen, C. T., Saputra, Y. M., Van Huynh, N., et al. (2020). *Enabling and emerging technologies for social distancing: A comprehensive survey-rob*.
- Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2020). *Monitoring covid-19 social distancing with person detection and tracking via fine-tuned YOLOv3 and deepsort techniques*. arXiv preprint arXiv:2005.01385.
- Hossain, M. S., Amin, S. U., Muhammad, G., & Al Sulaiman, M. (2019). Applying Deep Learning for Epilepsy Seizure Detection and Brain Mapping Visualization. *ACM Trans. Multimedia Comput. Commun. Appl. (ACM TOMM)*, 15(No. 1s), 1–17. Article 10.
- Punn, N. S., & Agarwal, S. (2020). Inception U-Net architecture for semantic segmentation to identify nuclei in microscopy cell images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(no. 1), 1–15.
- Punn, N. S., & Agarwal, S. (2019). Crowd analysis for congestion control early warning system on foot over bridge. in: *Twelfth IEEE International Conference on Contemporary Computing (IC3)*, 1–6.

- Qian, S., Zhang, T., Xu, C., & Hossain, M. S. (2015). Social Event Classification via Boosted Multimodal Supervised Latent Dirichlet Allocation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 11(2), 1–22. Article 27 (December 2014).
- Rahman, M. A., Zaman, N., Asyhari, A. T., Al-Turjman, F., et al. (2020). Data-driven dynamic clustering framework for mitigating the adverse economic impact of COVID-19 lockdown practices. *Sustainable Cities and Society*, 62, Article 102372.
- Redmon, J., & Farhadi, A. (2017). Yolo9000: better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7263–7271.
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. in *Advances in Neural Information Processing Systems*, 91–99.
- Rezaei, M., & Azarmi, M. (2020). DeepSOCIAL: Social distancing monitoring and infection risk assessment in COVID-19 pandemic. *medRxiv preprint*. <https://doi.org/10.1101/2020.08.27.20183277>
- Robakowska, M., Tyranska-Fobke, A., Nowak, J., Slezak, D., et al. (2017). The use of drones during mass events. *Disaster and Emergency Medicine Journal*, 2(no. 3), 129–134.
- Sathyamoorthy, A. J., Patel, U., Savle, Y. A., Paul, M., & Manocha, D. (2020). *COVID-Robot: Monitoring social distancing constraints in crowded scenarios*. arXiv preprint arXiv:2008.06585v2.
- Silva, B. N., Khan, M., & Han, K. (2018). Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. *Sustainable Cities and Society*, 38, 697–713.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. in: *Proc. of the 3rd International Conference on Learning Representations*. ICLR 2015, San Diego, CA, USA, May 7-9.
- Yang, X., Zhang, T., Xu, C., & Hossain, M. S. (2015). Automatic Visual Concept Learning for Social Event Understanding. *IEEE Transactions on Multimedia*, 17(no. 3), 346–358.
- Yang, X., et al. (2016). Deep Relative Attributes. *IEEE Transactions on Multimedia*, 18(no. 9), 1832–1842.
- Smart America (2020). URL: <https://smartamerica.org/> (Accessed on:04/11/2020).
- Sun, C., & Zha, Z. (2020). The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission. *Sustainable Cities and Society*, 62, Article 102390.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.
- Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., et al. (2018). Tensor2Tensor for neural machine translation. in: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, 193–199.
- WHO (2020). Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). World Health Organization. 30 January 2020. Archived from the original on 31 January 2020. Retrieved Aug 10, 2020.
- Yassine, Y., & Hossain, M. S. (2020). COVID -19 networking demand: An auction-based mechanism for automated selection of edge computing services. *IEEE Transactions on Network Science and Engineering*. <https://doi.org/10.1109/TNSE.2020.3026637>