COVID-19 scrutiny of social distancing using calibrate YOLO v3 and tracking & person detection alongside and Deepsort algorithm

Theerumurrthy Medasani
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu

Abstract— As of May 4, 2020, there had been around 3,519,901 real cases of this widespread disease (COVID-19), 247,630 fatalities worldwide, and the sickness had spread to more than 180 nations. The public is more exposed since there are no effective treatment options and no protection against COVID19. The only practical method of combating this epidemic because there are no vaccinations is through social distance. This idea serves as the inspiration for the essay, which suggests an in-depth study-based architecture to automate the process of utilizing captured footage to monitor social distance. In the suggested framework, persons are separated from background objects using YOLO v3 object detection model, and then tracked using the Deepsort technique, bounding boxes, and issued IDs. Further comparisons between the YOLO v3 model's output and other well-known state-of-the-art models are made in MAP terms and FPS, and object classification and localization determine loss values, such as SSD single shot detector and faster region-based convolution neural network (CNN). Following that, the bounding box dimensions and centroid coordinates are used to create a three-dimensional feature space from which the pairwise vectorized L2 norm is generated. To quantify not following social distance protocol, the violation index term is proposed. The findings of the experimental study depicts that the best outcomes are produced from the YOLO v3 with Deepsort tracking method, to measure social distance in real-time with a balanced mAP and FPS score.

Index Terms— Object tracing. Video monitoring, Object identification, social distancing, COVID-19

I. INTRODUCTION

COVID-19 COVID- 19 is a member of the group of illnesses brought on by corona viruses that were first identified in late December 2020 in the country CHINA, Wuhan. The WHO proclaimed this to be a epidemic disease dated March 11 [1], [2] post that it spread to 114 nations and resulted in 4000 fatalities and 118,000 active cases. The number of illnesses and fatalities recorded globally as of May 4, 2020, was over 3,519,901. The development of effective drugs and vaccines for this fatal virus is being pursued by several healthcare organisations, medical professionals, and scientists, but there has been no recorded progress to yet. In order to stem the spread of this contagious illness, the global society is under pressure to find alternative solutions. As per ongoing situations, social distancing is treated as the perfect solution to stop the disease spreading, also all the countries across the world followed social distancing to curb the spreading of

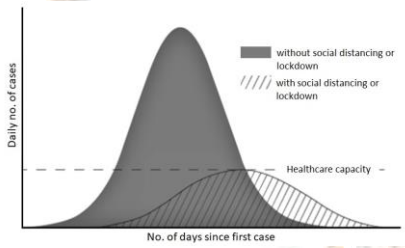


Fig. 1: Once the epidemic's peak was lowered and matched with the capacity of the healthcare system, social distance resulted.

disease accompanied by low loss of cost-effective efforts, and propose a resolution to perceive public gathering places distancing socially.

To reduce or stop the COVID-19 transmission, "social distance" is the ideal term to describe the direction of efforts. It seeks to limit interaction physically among infectious persons and non-infectious individuals. It is recommended that public adhere a minimal 6 feet distance among them as per the guidelines from world health organization [3], in order to amend social distancing.

Recent study says, maintaining social distance is crucial to preventing the spread of SARSCoV-2 as individuals with minimal or zero symptoms might unintentionally harbour the virus and infect others [4]. According to Fig. 1, maintaining appropriate social distance is the best strategy to minimise potentially contagious physical contact, which lowers the rate of infection [5], [6]. The lower peak might undoubtedly be compatible by the current health care setup and assist in providing patients fighting the coronavirus pandemic with better facilities. The analysis of variables and causes behind the infectious illnesses widespread is known as epidemiology. Scientific models are almost often the first choice for researching epidemiologic scenarios. The ancestor of nearly all models is classic SIR model of Kermack and McKendrick, which was initially created in 1927 [7]. Several studies of speculative widespread models and organic systems were conducted due to the model SIR and its developments by the acceptance structure [8], which has been the subject of numerous research publications.

While treating or finding strategies to control the spread of infectious respiratory disorders in the community, it is crucial to rate evaluation and route of spread of the causative disease. There is currently no well-known therapy that can be used to treat COVID-19, despite the efforts of multiple health organisations and epidemic scientists to discover the vaccinations. As a result, everyone in the globe takes precautions to limit the spread of sickness. Recent research by Eksin et al. [8] introduced a model called SIR which was modified along with distancing socially constraint, a (I,R), that is considered by means of

the numbers affected and improved individuals, denoted as I and R, correspondingly.

$$\frac{dS}{dt} = -\beta S\frac{I}{N}a(I,N)$$
$$\frac{dI}{dt} = -\delta I + \beta I\frac{I}{N}a(I,N)$$
$$\frac{dR}{dt} = \delta I$$

(1)

where δ depicts rate of recovery *and* β depicts the rate of infection. The size of population is calculated as $N = S + I + R$. The term social distance here ($a(I,R) : \mathbb{R}^2 \epsilon [0,1]$) links the rate of transition to an infected state (I), from a susceptible state (S) that is evaluated by $\frac{a\beta SI}{N}$.

There are different kinds of social distancing models and the initial classification is known as "long-term awareness", the communication existence among people is depreciated uniformly through the collective proportion of infected (Recovered and infected) individuals (Eq. B),

$$a = \left(1 - \frac{I+R}{N}\right)^k \quad \text{(B)}$$

In the Eq. C, it shows the later model called "short-term awareness", where the percentage of infected people at a particular instance is inversely correlated with the drop in contact.

$$a = \left(1 - \frac{I}{N}\right)^k \quad \text{(C)}$$

The behaviour parameter is denoted by k as, $k \geq 0$. Increased value of k shows the persons getting delicate towards the illness occurrence.

A company named Landing AI headed by Dr. Andrew Ng [10, 11], one of the most well-known names in artificial intelligence, declared the AI tool creation to track the distancing socially at work on April 16, 2020. The company claimed in a brief blog post that a potential solution could assess live video signals from the camera to figure out whether people are maintaining a secure physical distance among them or not. Also mentioned that this technology can easily relate to the present security cameras that are already present at various businesses in order to sustain a harmless distance among the workers. A short demo which was made available to show how to monitor social distance outlines three steps: calibration, detection, and measurement. Gartner, Inc. named AI Landing as one of the Cool Vendors in Core Technologies of AI on April 21, 2020, in honour of their in-time effort on this cutting-edge field to aid the struggle with the COVID -19 [12].

As a result, the authors of the current work were motivated to evaluate and compare how well-known object identification and tracking systems tracked social distance. The remaining paper sections are grouped as: The state-of-the-art object identification & trailing replicas are presented under III Section after Section II covers up most recent research proposals in this area of study. A deep learning-based system is later suggested in Section IV to track social distance. The experimentation is detailed in Section V, along with the associated results, and the conclusion is given in Part VI. The future scope and problems are covered in Section VII, and Section VIII concludes the current research project.

II. STUDY of BACKGROUND AND ITS INTERRELATED EFFORT

When COVID-19 initially appeared in the China's Wuhan, in December 2019, a decision was taken to implement it on January 23, 2020, as an unprecedented step [13], as isolating socially is undoubtedly the utmost effective means to prevent the spread of contagious illnesses. The epidemic in China peaked in the first week of February, when the cases were in range of 2,000 and 4,000 daily confirmed cases newly, within a month. Finally, up until March 23, 2020, there had been no additional confirmed instances for five days in a row for the first time since the outbreak [14]. This demonstrates how social isolation policies, initially put in place in China to fight COVID-19, later spread throughout the world.

Prem et al.'s [15] objective was to identify the social isolation impact laws on the COVID-19 epidemic. The ongoing course of the outbreak was simulated by authors utilising susceptible-exposed-infected-removed (SEIR) models with artificial location-specific interaction patterns. Moreover, it was argued that easing social restrictions prematurely and abruptly would cause a early secondary peak, that might be smoothed by easing the intrusions slowly [15]. We all know that the best way to flatten the infection curve is to practise social isolation, which is monetarily costly yet necessary. Adolph et al. [16] drew attention to the predicament facing the USA, where the absence of consensus among politicians prevented early adoption, continuing public health consequences. Nonetheless, there is a trade-off between the region's economic standing and how rigid social distance is. The research suggests that modest points of activity might be abided while averting an outbreak that spreads widely.

Several countries had used solutions of technology-based in different volumes to limit the new coronavirus pandemic ever since it started [18], [19], [20]. Several sophisticated nations, such as India and South Korea, for example, using GPS to track the whereabouts of suspects or diseased people in order to keep an eye on any potential for exposure of healthy people. The Indian government makes use of the Mobile Application called Arogya Setu to identify COVID-19 affected individuals in the nearby premises [21]. Nonetheless, some law enforcement agencies have started deploying drones and other types of monitoring equipment to see large crowds and then take appropriate measures to scatter them [22], [23]. Those physical involvement in these tough conditions may aid to level the curve, as well as it entails hazards for the general public and challenges for the workforce.

Even though it relies on manual techniques to spot odd activity, detection of humans utilising video surveillance systems is a well-known part of research [24]. Nevertheless, it has several limitations. The necessity for intelligent systems to recognise and record human activity is supported by recent breakthroughs in this regard. Given several limitations, including low-resolution video, different articulated poses, clothing, lighting, and complex backgrounds, in addition to limited machine vision capabilities, the goal of human detection is challenging [25]. Nonetheless, recognising these limits can aid with detection performance.

The initial two processes in identifying an object in motion are detecting the objects and classification of objects [26, 27]. During the first stage of detecting the objects, optical flow [29], background removal [28] and spaciotemporal filtering methods [30] may be helpful. The background subtraction approach calculates the difference at the block level or pixel level amongst the present frame

and a background frame (first frame) [31]. The most popular techniques for background removal are warping background, non-parametric background, hierarchical background models, temporal differencing, adaptive Gaussian mixture, and warping background [32]. The optical flow-based identification of objects technique [29], which characterises vectors flow connected to the item's motion through time [33], may be used to recognise moving objects on a set of pictures. Researchers claim that flow-based optical techniques consist of calculative costs and remain vulnerable to several gesture-related irregularities, such as, lighting, noise, colour etc. [34]. In Aslani et alfilter-based technique for motion detection [30], to determine the motion parameters the 3D spatial-temporal features of the subject affecting in the image stream are used. The performance of above techniques is limited by noise and altering pattern uncertainty, despite the simplicity and lack of processing complexity that make them appealing [35].

Recent breakthroughs in sophisticated approaches have effectively handled object detecting issues. In the past ten years, region proposal techniques have been used by convolutional neural networks (CNN), faster region-based CNN, and region-based CNN to generate object scores prior to classification and to create bounding boxes around the object of visualisation interest and additional numerical study [38]. These techniques are efficacious, but are required more instruction time [39]. A regression-based method different approach known as YOLO considers to determine the class chances and to size the boxes of bounding and inside of them given that all these CNN-based algorithms sort data [40]. With the class probability ratings for each component being taken into consideration as an item, this method successfully divides the image into several bounding box-representing portions. This method gives substantial speed gains while giving up efficiency in exchange for speed. Powerful generalisation capabilities for encoding a whole picture are demonstrated by the detector module [41].

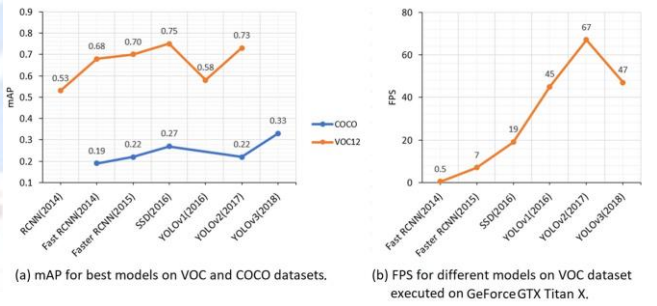
Many study findings have been published in the past several years based on these above notions. With various social implications, crowd counting has emerged as a promising study field. Eshel et al[42] .'s work on crowd recognition and individual counting proposed several height homographs for detection of head top and addressed the obstructions issue in applications related to surveillance of video. Based on the idea of crowd counting, Chen et al. [43] created a promotion application which is electronic. A public counting model which was vision-based presented by Chih-Wen et al. [44] for a related application. After that, Yao et al. [45] produced contributions from fixed photographic camera to perform background removal in order to shape of the crowd in films and to shape the model for the appearance.

Based on form, motion-dependant or texture-based data, classification algorithms may be used to find a person if an object has been discovered. Shape-based approaches use moving areas like points, boxes, and blobs to gather shape-related information that is used to identify the person. This technique works badly because of various limitations in popular template-matching systems [46, 47]. Using a part-based template matching method improves performance even better [48]. Dalal et al. [49] proposed texture-based methods for people detection, like histograms of oriented gradient (HOG), that employ support vector machine (SVM) and high-dimensional features based on edges.

Using face [50], [51] and recognition of gait [52] approaches, recent research has shown that additional identification of a person through video surveillance is possible. Unfortunately, due to partial or complete occlusion issues, it might be challenging to recognise and track persons in a crowd at times. While Leibe et al. [53] proposed a resolution based on path estimation, Andriluka et al. [54] anticipated a method to find people who are partially obscured using tracklet-based detectors. Yilmaz et al. [55] review a wide range of additional tracking methods, as well as various object and motion representations.

In the area of video surveillance, several studies have been conducted. The 6 different types of motions are shown by KTH human motion dataset [56], whereas the 11 different types of activities are shown by the INRIA XMAS multi-view dataset [57] and 10 different types of activities are shown human action dataset by the Weizmann [58]. The distinct dataset that was formed by a team of academics from Oxford University [59] is called as PETS-Performance Evaluation of Tracking and Surveillance. This is used in visualization-based study and contains several separate datasets for various computer vision tasks. In the current work, Open image datasets [60] are taken into consideration in order to optimise the object identification and trailing algorithms for detecting the human. The models are taught to recognise individuals from a pool of 19,957 classifications. The bounding boxes that represent the person's image are labelled at the picture-level, and the accompanying coordinates are also provided. The refined suggested system is also used to recreate the Oxford town centre footage of surveillance [23] in order to track communal estrangement.

A unique data model with consistent observations for object recognition, multimodal image descriptions, visual relationship detection, image classification, instance segmentation will make it easier to effectively know and perform object recognition tasks, and further the evolution toward a true knowledge of the section. These works which was reviewed and corresponding research projects evidently indicate the recognition of human use and can be extended to a variety of conditions to have the present requirements, like verifying established standards for social distance, work practises and hygiene etc.



(a) mAP for best models on VOC and COCO datasets.

(b) FPS for different models on VOC dataset executed on GeForce GTX Titan X.

Fig. 2: Most common object identification models' performance overview on the MS-COCO datasets and PASCAL-VOC.

III. MODELS OF TRACKING AND OBJECT DETECTION

In the Fig. 2 shown, the effective object identification models tested on the MS-COCO [67] and PASCAL-VOC [66] datasets, such as the fast RCNN [62], RCNN [61], faster RCNN [38], YOLO v2 [64], SSD [63], YOLO v1 [40], and YOLO v3 [65], display trade-off between perfection and speed of the detection, which depends on a variety of aspects. A feature separator inclines to encrypt the input from the model into a particular feature depiction, that helps in the study and detecting the ways connected to the desired items. For the ILSVRC

ImageNet challenge [71], Table I compiles the accuracy performance of each of these well-known and efficient feature abstraction networks, with training duration and speed directly impacted by the number of trainable parameters. As indicated in Table I, Inception v2 was able to attain a sufficient classification precision with a limited amount of trainable parameters. It is therefore used as the core architecture for the SSD object and faster RCNN recognition models, enabling more precise computations. As indicated by Redmon et al. [65], YOLO v3 employs a diversified architecture, Darknet-53, as opposed to Inception v2's.

A. Anchor boxes

in a scene [36] a careful analysis of the literature revealed that anchor boxes are a characteristic shared by all well-known object recognition models that are used to identify a different object. Across several geographic places and range in size and aspect ratio (per filter), the boxes are placed over the input image.

TABLE I: Performance issues with the ImageNet extraction of feature network.

Backbone model	VGG-16 [68]	Resnet v2 [72]	ResNet-101 [69]	Inception v3 [72]	Inception v2 [70]
Accuracy (a)	0.71	0.8	0.76	0.78	0.74
Parameters (p)	15 M	54 M	42.5 M	22M	10 M
Ratio (a*100/p)	**4.73**	**1.48**	**1.78**	**3.58**	**7.4**

TABLE II: Anchor boxes Generation the with hyperparameters.

Model of Detection	Aspect ratio (r)	Vector Size (p)	IoU th. for NMS	Anchor boxes
Faster RCNN	[0.5, 1.0, 2.0]	[0.25, 0.5, 1.0]	0.7	9
SSD	[0.3, 0.5, 1.0]	[0.2, 0.57, 0.95]	0.6	9
YOLO v3	[0.5, 1.0, 2.0]	[0.25, 0.5, 1.0]	0.7	9

Take the constraints, size as $p \in (0, 1]$ and characteristic ratio as $r > 0$, then with dimensions as $bp\,r \times hp\,r$, for a certain location of the anchor boxes in an $\sqrt{\ }\sqrt{\ }$ image which can be created.

A three-dimensional location includes multiple anchor boxes; therefore, an item may relate to more than one of them. This issue is resolved by using the intersection over union (IoU) field, which restricts the anchor boxes' affiliation through the item of attention. By dividing the number of areas that overlap among the allocated box of anchor and the truth which is grounded by the sum of those areas, the score is determined. The box which is best of bounding for an element is then identified with the comparison with the provided limited hyper parameter of the score value. Table II demonstrates p and r settings for each model.

1)Function of loss: For every level of training type, anchor box which was predicted "*a*" is allotted a negative (0) and positive (1) or label depending on relationship to the interest object containing ground-truth box "*t*." Following that, the class label for the positive anchor box is formed, with $z_o = 0$ for negative anchor boxes and c_n displays the category of the n^{th} item. Consider the following case: According to Eq. 4, the loss for a single anchor forecast (L_{cls}) and the bounding box regression loss (L_{reg}), may be calculated for an image 'P' after a training parameters model 'k' forecasted the class of object as $Z_{cls}(P|a;k)$ and the associated box as $Z_{reg}(P|a;k)$.

$$L(a|P;k) = \alpha.1^{obj}{}_a\, L_{reg}(f(t_a|a) - Z_{reg}(P|a;k)) + \beta.L_{cls}(z_a, Z_{cls}(P|a;k)) \quad (4)$$

where $1^{obj}{}_a$ is 1 if 'a' is a anchor which is positive, β *and* α are the associated weights with the regression and sorting loss. Later, the computation of complete loss model as

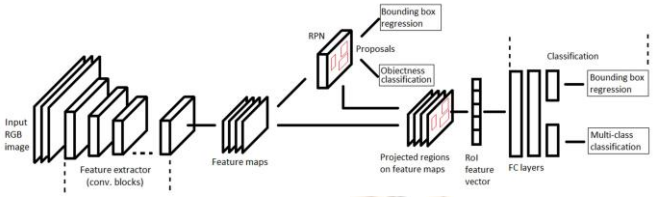


Fig. 3: Architecture Faster RCNN

the $L(a|P;k)$ average on top of the forecasts for all the anchors.

B. Faster RCNN

The RCNN and fast RCNN, which both depend on the external region proposal technique based on selective search (SS) [73], gave rise to the faster RCNN [38]. The advantages of convolution layers are recommended rather than the SS for better and faster object localisation, according to many academics [74–76]. The RPN which uses CNN models like ResNet.VGGNet, etc. to make faster RCNN 10 times speeder than fast RCNN, was proposed to generate region proposals. The faster RCNN architecture depicted in Fig. 3 is schematic; it has a Region Proposal Network module that makes binary sorting of an object or the background, and a classification module that performs multiclass classification on extracted feature maps with the help of region of interest (RoI) pooling [38] with projected regions to assign categories to each detected object.

*1) Function of loss: B*y combining the fast RCNN detector and the RPN module, he faster RCNN is created. The loss of classification and box bounding loss of regression described in Eq. 4 along with the functions L_{cls} and L_{reg} stated in Eq. 5 make up the overall multi-task loss function.

$$L_1 \quad (q) = \begin{cases} 0.5q^2, & if\ |\ q\ |< \\ |\ q\ | - 0.5, & \text{otherwise} \end{cases}$$
$$L_{cls}(p_i, p_i^*) = -p_i^* \log(p_i) - (1 - p_i^*) \log(1 - p_i)$$
$$L_{reg}(t^u, v) = \sum_{x \epsilon x, y, w, h} L_1^{smooth}(t_i^u - v) \quad (5)$$

smooth 1.

.

Here the forecasted alterations t^u of the bounding box $t^u = \{t_x^u, t_y^u, t_w^u, t_h^u\}$. Here u is a genuine class label with width w, v is a ground-truth bounding box, $p^*{}_i$ is the predicted class and p_i is the actual class, (x, y) corresponds to the top-left coordinates of the bounding box and with height h.

C. Single Shot Detector (SSD)

In this study for an additional detection of objects technique to find individuals in an actual-time surveillance video system SSD is used. According to what was previously discussed, faster R-CNN works on area suggests to construct boundary boxes to denote objects, demonstrates better accuracy, but FPS. Using multiscale structures and defaulting boxes in a solo process, SSD enhances precision and FPS for real-time processing.

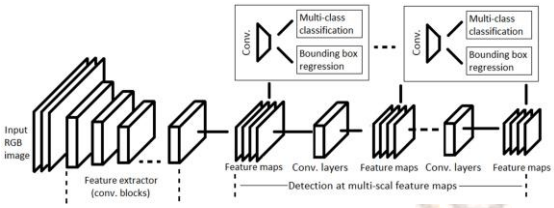


Fig. 4: SSD architecture

The network creates boxes of bounding defined dimensions and a count constructed in those boxes on the existence of instances object class after the feed-forward convolution, the final detections are done using the NMS step. There are two stages involved in using a three-part architecture to identify objects: Feature map extraction and convolution filter application. In the first portion, feature maps are extracted using a base pretrained network, and in the second, multiscale feature layers are employed with a cascade of convolution filters. The last component, a in-extreme compressive device, gets rid of overlapping boxes and only permits one item per box as seen in Fig. 4.

1) Loss function: The SSD model's overall loss function is equals total amount of the multi-class sorting loss (L_{cls}) and box bounding loss of regression (localization loss, L_{reg}), like the faster RCNN model that was previously described, where L_{reg} and L_{cls} are defined by the following equations

$$L_{reg}(x,l,g) = \sum_{i\epsilon pos}^{N} \sum_{m\epsilon c_x,c_y,w,h} x_{ij}^{k} smooth_{L_1}(l_i^m - \hat{g}_j^m) \quad ,$$

$$\hat{g}_j^{c_x} = \frac{(g_j^{c_x} - a_i^{c_x})}{a_i^w}, \ \hat{g}_j^{c_y} = \frac{(g_j^{c_y} - a_i^{c_y})}{a_i^h}, \quad (6)$$

$$\hat{g}_j^w = \log\left(\frac{g_j^w}{a_i^w}\right), \ \hat{g}_j^h = \log\left(\frac{g_j^h}{a_i^h}\right),$$

$$x_{ij}^p = \begin{cases} 1, & \text{if IoU} > 0.5 \\ 0 & \text{otherwise.} \end{cases} ,$$

Where g is the ground truth box, l is the predicted box, c_x and c_y are offsets to the anchor box a, x^p_{ij} is an indicator that matches the i^{th} anchor box to the j^{th} ground truth box.

$$L_{cls}(x,c) = -\sum_{i\epsilon Pos}^{N} x_{ij}^p \log(\hat{c}_i^p) - \sum_{i\epsilon Neg} \log(\hat{c}_i^o) \quad (7)$$

where $\hat{c}_i^p = \frac{\exp c_i^p}{\sum_p \exp c_i^p}$ and N is the number of default matched boxes.

D. YOLO

YOLO [40] is another rival of SSD for object recognition. By only taking a single glance at the image, this technique shows the kind and object location. Instead of classifying the object detection issue, YOLO assigns class possibilities to the anchor boxes as a regression task.

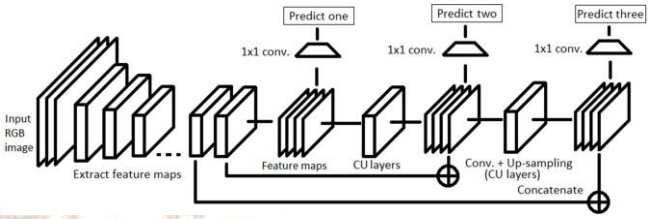


Fig. 5: YOLO v3 Graphic illustration

Many boxes of bounding and probabilities of class are concurrently forecasted through a unique convolutional network. YOLO comes in three main iterations: v1, v2, and v3. Google Net (Inception network), which is intended for categorization of object in images, served as the model for YOLO v1. These only services a drop layer, followed by layers of convolutional, as opposed to the modules of Inception utilized by GoogleNet. The YOLO v2 [64] goal is to greatly increase accuracy while speeding up the process. With 19 convolutional layers, an output softmax layer & five max layers of pooling for categorization of object, the backbone network Darknet-19, of YOLO v2, is used. With notable gains in FPS, mAP, and classification of object score, YOLO v2 beat its forerunner (YOLO v1). On the other hand, YOLO v3 conducts multi-label sorting using different classifiers as opposed to SoftMax as in the case of YOLO v1 and v2. As a skeleton planning for YOLO v3 that abstracts feature classification maps, Redmon et al. suggested Darknet-53. Remaining blocks (short connections) and up sampling levels for concatenation and extra network depth make up Darknet-53 as opposed to Darknet-19. The solution to the problem of ineffectively identifying small objects is YOLO v3, which creates 3 forecasts for every three-dimensional position at various scales in an image [77]. Computation of objectness, boundary box regressor, and classification scores makes it feasible to keep track of each prediction. In Figure 5, the YOLOv3 planning is shown schematically.

1)Function of Loss: The YOLO v3 total function of loss is composed of cross entropy, localization loss and confidence loss for categorization score.

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{obji,j} \left[(t_x - \hat{t_x})^2 + (t_y - \hat{t_y})^2 + (t_w - \hat{t_w})^2 + (t_h - \hat{t_h})^2 \right]$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1^{obj}_{i,j} \left[-\log(\sigma(t_o)) + \sum_{k=1}^{C} BCE(\hat{y_k}, \sigma(s_k)) \right]$$

$$+ \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{noobji,j} \left[-\log(1 - \sigma(t_o)) \right] \quad (8)$$

where λ_{coord} denotes the importance of the error coordinates, S^2 denotes grids quantity as in picture, and boxes of bounding count made per grid is denoted by B . $1_{i,j}^{obj} = 1$ defines that entity restrictions in the j^{th} box of bounding otherwise it is 0. in grid i,

E. Deepsort

A Deepsort method which is deep learning-based is used in the current study to track people seen in the surveillance video [78]. In order to predict the corresponding paths of the interest objects, it uses patterns acquired through the detection of objects in the images, which are subsequently combined with historical data. By assigning distinctive identifiers to each object under consideration, it maintains track of them for later statistical analysis. Deepsort is also helpful for dealing with related issues like numerous perspectives, labelling training data, occlusion, and non-stationary recording devices. For effective tracking, the Kalman filter and Hungarian algorithm are used. When used recursively, the Kalman filter improves association by forecasting future locations based on the current position. To identify if an object in the present setting is equivalent as an object in the earlier frame for association and id attribution, a Hungarian method is used. After a faster RCNN is set up for person recognition, the following defines each object in an eight-dimensional space using a linear constant velocity model is shown [79]:

$$x = [r,s,\lambda,h,p_r,q_r,\lambda_r,h_r]^T \quad (9)$$

Here, h is the height of the image, (r,s) is the centroid of the box bounding, a is the aspect ratio and the other variables are the respective velocities of the variables. The coordinates of bounding (r,s,λ,h) are considered as straight remarks of the state of object with velocity motion constant and observation model linear later in the standard Kalman filter.

The entire quantity of frames is prepared for each path k, originally from the previous positive measurement connection a_k. The quantity is increased upon a successful prediction, and it is later reset to zero when the track is linked to a dimension. If the discovered tracks are older than a predefined maximum age, indicating that the related objects have left the section, the related track is also deleted from the collection of tracks. For each unexplained track of recently discovered things which could not be plotted to the prevailing tracks, new track hypotheses are also formed if there are zero tracks accessible for spotted items. The new tracks are categorised as uncertain for the first three frames pending the creation of an efficient measurement mapping. If measurement efforts to map them are unsuccessful, the tracks are eliminated from the track collection. Then, using the Mahalanobis distance obtained between motion and appearance information as described in Eq. 10, the mapping issue between the recently received data and the anticipated Kalman states is resolved.

$$d^{(1)}(i,j) = (d_j - y_i)^T S_i^{-1}(d_j - y_i) \quad (10)$$

where the forecast of the i^{th} path spreading into measurement space is denoted by (y_i,S_i) and the j^{th} box of bounding recognition by d_j. By estimating the number of standard deviations, the Mahalanobis distance takes this uncertainty into account when detecting deviations from the position of the mean track. Additionally, by thresholding the Mahalanobis distance when using this measure, it is possible to rule out associations that are unlikely. This choice is represented with a pointer that assesses to 1 if the connotation between the i^{th} track and j^{th} recognition is admissible (Eq. 11).

$$b_{i,j}^{(1)} = 1[d^{(1)}(i,j) < t^{(1)}] \quad (11)$$

Despite its effectiveness, Mahalanobis distance fails in situations where possibility of camera motion; as a result, additional measure is presented for the projected issue. The smallest cosine distance between the i^{th} track and j^{th} detection in appearance space is measured by this second metric as follows:

$$d^{(2)}(i,j) = min\{1 - r_j{}^T r_k{}^{(i)} \mid r_k{}^{(i)} \in \mathbb{R}^2\} \quad (12)$$

Once more, a binary variable is used to show whether an association is valid in light of the following metric:

$$b_{i,j}^{(1)} = 1[d^{(2)}(i,j) < t^{(2)}] \quad (13)$$

Together metrics are collective with a prejudiced sum to create the combined contention:

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1 - \lambda)d^{(2)}(i,j) \quad (14)$$

if an association falls within the gating area of both metrics, it is acceptable:

$$b_{i,j} = \prod_{m=1}^{2} b_{(i,jm)}. \quad (15)$$

Through the use of hyperparameters λ, each metric's impact on the total connection cost can be managed.

IV. Proposed Approach

Deep learning's development has introduced the top performance methods for a range of tasks and problems, including speech recognition [76], machine translation [75], and medical diagnosis [74]. Most of these activities revolve around classifying, detecting, segmenting, tracking, and recognizing objects [81], [82]. As can be seen in Fig. 2, which shows the performance of such models compared to mAP and FPS on the popular benchmark datasets PASCAL-VOC [66] and MS-COCO [67] and their associated hardware resources, convolutional neural network (CNN)-based architectures have shown significant performance improvements leading to high-quality object recognition.

In the current study, a Deepsort framework is anticipated to assist with the social distance remedy for managing the increase of COVID-19 situations. This system uses object identification and tracking models. YOLO v3 [65] and Deepsort [78] are utilised as detecting the objects and tracking techniques, and every identified item is ringed by bounding boxes in order to retain the balance between speed and precision. Afterwards, these boxes of bounding are applied to calculate the pairwise L2 norm with calculatedly efficient vectorized depiction for recognising the clusters of persons not respecting the order of distancing socially. Additionally, to see the clusters in the live stream, every box of bounding is color-coded depending on its affiliation with the group where members belonging to the same group are shown with the similar colour. A plot which is streamlined

shows the analytical breakdown of the quantity of public organizations is also included in each surveillance frame, along with an index word that shows the ratio of individuals to groups. The expected count of desecrations may be determined by multiplying the total social groupings with the score of violation.

A. Workflow

The crucial stages made to create a context for tracking social distance are included in this section.

1. To accurately track and find the individual from a video, fine-tune the skilled object detection method.
2. There are three-dimensional (l,s,t), features associated with each individual, where the coordinates centroid of the box of bounding are defined as (l, s) and d defines the individual depth as observed from the camera [83].

$$t = ((2 * \pi * 180)/(width + height * 360) * 1000 + 3) \quad (16)$$

3. The surveillance film is fed to the trained model. For respective recognized individual, the model creates a collection of boxes of bounding and an ID.
4. Pairwise *L2* norm is computed for Bounding boxes set, as given by the subsequent calculation.

$$||D||_2 = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (17)$$

In which work $n = 3.5$.

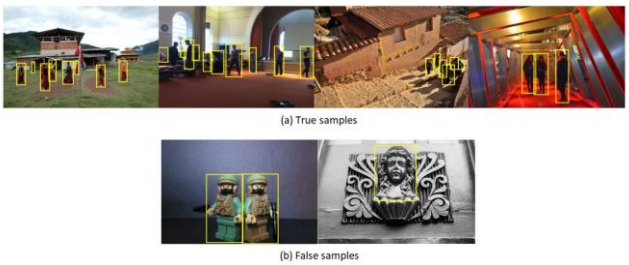
5. Neighbours for each person who meets the closeness sensitivity are then assigned using the dense matrix of *L2* norm. With numerous trials, the three-dimensional position of the individual in a setting with a range of (90,170) pixels are used to dynamically update the closeness threshold. Everyone that matches the proximity stuff is given a neighbour or neighbours forming a set depicted in various coding colours in contrary to other individuals.
6. The emergence of organizations suggests that social distance is no longer practiced, and the following metrics are used to quantify this:

Consider m_g as amount of groups or clusters identified, and m_p as total amount of people found in close proximity.

$I_i = m_p/m_g$, where I_i is the index of violation.

V. EXPERIMENTS AND RESULTS

The dataset used in the object recognition models described above was maintained by Google's open-source community and obtained from the Open Image Dataset (OID) repository [73], and was fine-tuned for binary classification (person or no person) using Inception v2 as the backbone network on the Nvidia GTX 1060 GPU



.

Fig.6: Data samples from the open image dataset demonstrating (a) actual illustrations and (b) fake illustrations of a "Person" class.

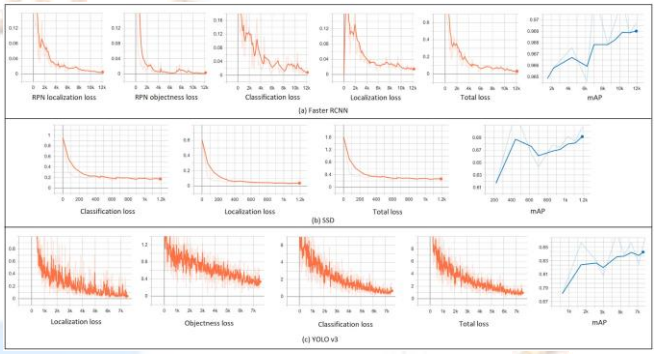


Fig. 7: Training phase: Losses for the object recognition models on the OID validation set per iteration

Through the OIDv4 toolkit [84], the various images with the class name "Person" are downloaded along with the annotations. The dataset, which consists of 800 images and was acquired by manually selecting only the true samples, is shown in Fig. 6 as samples. The dataset is then split in half, 8:2, into training and testing groups. The testing collection also includes frames from surveillance video taken in the Oxford town center to further strengthen the testing [23]. Subsequently, this video is also used to model the general strategy for keeping an eye on social distance. This video is then used to demonstrate the basic approach for monitoring the social distance. And mAP, together with the sorting, and total loss in the identification of the individual, localization is all utilised to continually check the models' performance during the training phase, as shown in Fig. Seven. 3rd table provides a summary of every model's finding at the training phase end, along with the total loss (TL), number of iterations (NoI) and mAP values. training time (TT). The faster RCNN model, which obtained the lowest loss with the highest mAP, has the lowest FPS, making it unsuitable for real-time applications, as is evident from the research. Additionally, with equal and FPS score, YOLO v3,mAP, training time, outperformed SSD in terms of performance. Then to track social distance on the surveillance footage, the trained YOLO v3 model is used.

TABLE III: Comparison of the detection of object algorithms' performance.

	Models		
	Faster RCNN	**SSD**	**YOLO v3**
TT (in sec.)	9651	2124	5659
NoI	12135	1200	7560
mAP	0.969	0.691	0.846
TL	0.02	0.22	0.87
FPS	3	10	23

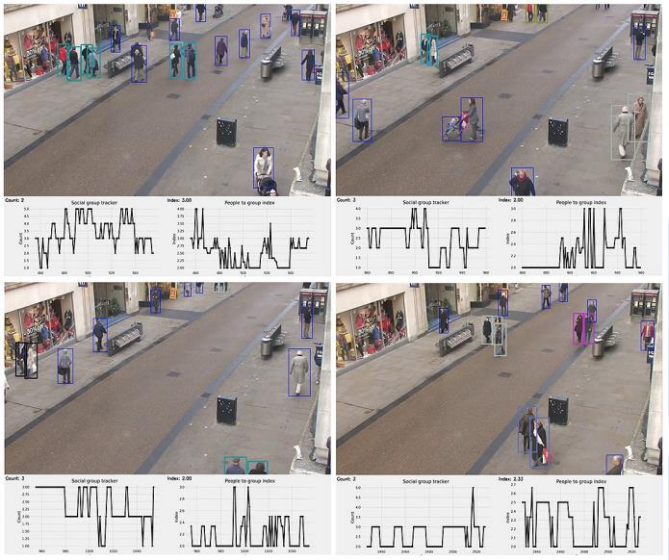


Fig. 8: Examples of the results from the proposed tracking social distance on the video surveillance system at Oxford Town Center.

VI. Output

The outcome of the recommended framework is illustrated in Fig. 8, which also replicates statistical analysis by providing various public sets represented by identical colour imbibing and the ratio of the set of persons to the set of groups computed by an index term of violation. The bounding boxes are used to contain the discovered individuals. The violation score for the frames in Fig. 8 is 3, 2, 2, and 2.33. The frames containing identified violations are timestamped and archived for future examination.

VII. Future scope and challenges

Accuracy and precision are crucial to the success of this application because it can be used in any workplace. An increase in false positives could make individuals being watched feel uneasy and anxious. Additional steps can be taken to address legitimate privacy and individual rights concerns, like obtaining former agreement for those occupied environments, generally hiding one's individuality, and sustaining opaqueness on its proper uses amongst a small set of stakeholders.

VIII. Conclusion

Bounding boxes are created to identify collections of individuals or clusters which satisfy closeness object determined by a vectorized pairwise method. By evaluating the set of collections created and using a index term violation (which is calculated as the ratio of the set of people to the set of groups), the violations count stand confirmed.

A deep learning-practical based was proposed in the paper, that uses bounding boxes to identify everyone in real-time. This aids in process automation of observing social distance using detection methods and tracking the items. In the extended trials, the most recent object recognition models SSD, faster YOLO v3, RCNN were used. The performance of YOLO v3 was efficient and had a balanced mAp and FPS result. This technique is enormously delicate to the spatial assignment of the vision of camera field.

Acknowledgment

The writers sincerely thank their peers for their insightful criticism and recommendations. The Department of Science and Technology (DST), Government of India (GoI) and the Interdisciplinary Cyber Physical Systems (ICPS) Programme provided financial assistance to the authors for conducting background research, which was crucial to the success of the current research work (Reference No. 244).

References

[1] W. H. Organization, "WHO corona-viruses (COVID-19)," https://www. who.int/emergencies/diseases/novel-corona-virus-2019, 2020, [Online; accessed May 02, 2020].

[2] WHO, "Who director-general's opening remarks at the media briefing on covid-19-11 march 2020." https://www.who.int/dg/speeches/detail/, 2020, [Online; accessed March 12, 2020].

[3] L. Hensley, "Social distancing is out, physical distancing is in—here's how to do it," *Global News–Canada (27 March 2020)*, 2020.

[4] ECDPC, "Considerations relating to social distancing measures in response to COVID-19 – second update," https://www.ecdc.europa.eu/ en/publications-data/considerations, 2020, [Online; accessed March 23, 2020].

[5] M. W. Fong, H. Gao, J. Y. Wong, J. Xiao, E. Y. Shiu, S. Ryu, and B. J. Cowling, "Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings—social distancing measures," 2020.

[6] F. Ahmed, N. Zviedrite, and A. Uzicanin, "Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review," *BMC public health*, vol. 18, no. 1, p. 518, 2018.

[7] W. O. Kermack and A. G. McKendrick, "Contributions to the mathematical theory of epidemics–i. 1927." 1991.

[8] C. Eksin, K. Paarporn, and J. S. Weitz, "Systematic biases in disease forecasting–the role of behavior change," *Epidemics*, vol. 27, pp. 96– 105, 2019.

[9] M. Zhao and H. Zhao, "Asymptotic behavior of global positive solution to a stochastic sir model incorporating media coverage," *Advances in Difference Equations*, vol. 2016, no. 1, pp. 1–17, 2016.

[10] P. Alto, "Landing AI Named an April 2020 Cool Vendor in the Gartner Cool Vendors in AI Core Technologies," https://www.yahoo.com/ lifestyle/landing-ai-named-april-2020-152100532.html, 2020, [Online; accessed April 21, 2020].

[11] A. Y. Ng, "Curriculum Vitae," https://ai.stanford.edu/~ang/ curriculum-vitae.pdf.

[12] L. AI, "Landing AI Named an April 2020 Cool Vendor in the Gartner Cool Vendors in AI Core Technologies," https://www.prnewswire.com/ news-releases/, 2020, [Online; accessed April 22, 2020].

[13] B. News, "China coronavirus: Lockdown measures rise across Hubei province," https://www.bbc.co.uk/news/world-asia-china51217455, 2020, [Online; accessed January 23, 2020].

[14] N. H. C. of the People's Republic of China, "Daily briefing on novel coronavirus cases in China," http://en.nhc.gov.cn/2020-03/20/c _78006. htm, 2020, [Online; accessed March 20, 2020].

[15] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, S. Flasche, S. Clifford, C. A. Pearson, J. D. Munday *et al.*, "The effect of control strategies to reduce social mixing on outcomes of the covid19 epidemic in wuhan, china: a modelling study," *The Lancet Public Health*, 2020.

[16] C. Adolph, K. Amano, B. Bang-Jensen, N. Fullman, and J. Wilkerson, "Pandemic politics: Timing state-level social distancing responses to covid-19," *medRxiv*, 2020.

[17] K. E. Ainslie, C. E. Walters, H. Fu, S. Bhatia, H. Wang, X. Xi, M. Baguelin, S. Bhatt, A. Boonyasiri, O. Boyd *et al.*, "Evidence of initial success for china exiting covid-19 social distancing policy after achieving containment," *Wellcome Open Research*, vol. 5, no. 81, p. 81, 2020.

[18] S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan, "Target specific mining of covid-19 scholarly articles using one-class approach," 2020.

[19] N. S. Punn and S. Agarwal, "Automated diagnosis of covid-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks," 2020.

[20] N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Covid-19 epidemic analysis using machine learning and deep learning algorithms," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/ early/2020/04/11/2020.04.08.20057679

[21] O. website of Indian Government, "Distribution of the novel coronavirus-infected pneumoni Aarogya Setu Mobile App," https:// www.mygov.in/aarogya-setu-app/, 2020.

[22] M. Robakowska, A. Tyranska-Fobke, J. Nowak, D. Slezak, P. Zuratynski, P. Robakowski, K. Nadolny, and J. R. Ładny, "The use of drones during mass events," *Disaster and Emergency Medicine Journal*, vol. 2, no. 3, pp. 129–134, 2017.

[23] J. Harvey, Adam. LaPlace. (2019) Megapixels.cc: Origins, ethics, and privacy implications of publicly available face recognition image datasets. [Online]. Available: https://megapixels.cc/

[24] N. Sulman, T. Sanocki, D. Goldgof, and R. Kasturi, "How effective is human video surveillance performance?" in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–3.

[25] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern recognition letters*, vol. 34, no. 1, pp. 3–19, 2013.

[26] K. A. Joshi and D. G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *International Journal of Soft Computing and Engineering*, vol. 2, no. 3, pp. 44–48, 2012.

[27] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," in *European Conference on Computer Vision*. Springer, 2002, pp. 343–357.

[28] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background¨ subtraction techniques for video surveillance," in *CVPR 2011*. IEEE, 2011, pp. 1937–1944.

[29] S. Aslani and H. Mahdavi-Nasab, "Optical flow based moving object detection and tracking for traffic surveillance," *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, vol. 7, no. 9, pp. 1252–1256, 2013.

[30] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition´ via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[31] M. Piccardi, "Background subtraction techniques: a review," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, vol. 4. IEEE, 2004, pp. 3099–3104.

[32] Y. Xu, J. Dong, B. Zhang, and D. Xu, "Background modeling methods in video analysis: A review and comparative evaluation," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 1, pp. 43–60, 2016.

[33] H. Tsutsui, J. Miura, and Y. Shirai, "Optical flow-based person tracking by multiple cameras," in *Conference Documentation International Conference on Multisensor Fusion and Integration for Intelligent Systems. MFI 2001 (Cat. No. 01TH8590)*. IEEE, 2001, pp. 91–96.

[34] A. Agarwal, S. Gupta, and D. K. Singh, "Review of optical flow technique for moving object detection," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 2016, pp. 409–413.

[35] S. A. Niyogi and E. H. Adelson, "Analyzing gait with spatiotemporal surfaces," in *Proceedings of 1994 IEEE Workshop on Motion of Nonrigid and Articulated Objects*. IEEE, 1994, pp. 64–69.

[36] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[39] X. Chen and A. Gupta, "An implementation of faster rcnn with study for region sampling," *arXiv preprint arXiv:1702.02138*, 2017.

[40] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779– 788.

[41] M. Putra, Z. Yussof, K. Lim, and S. Salim, "Convolutional neural network for person and car detection using yolo framework,"

Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1-7, pp. 67–71, 2018.

[42] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[43] D.-Y. Chen, C.-W. Su, Y.-C. Zeng, S.-W. Sun, W.-R. Lai, and H.Y. M. Liao, "An online people counting system for electronic advertising machines," in *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 2009, pp. 1262–1265.

[44] C.-W. Su, H.-Y. M. Liao, and H.-R. Tyan, "A vision-based people counting approach based on the symmetry measure," in *2009 IEEE International Symposium on Circuits and Systems*. IEEE, 2009, pp. 2617–2620.

[45] J. Yao and J.-M. Odobez, "Fast human detection from joint appearance and foreground feature subset covariances," *Computer Vision and Image Understanding*, vol. 115, no. 10, pp. 1414–1426, 2011.

[46] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[47] F. Z. Eishita, A. Rahman, S. A. Azad, and A. Rahman, "Occlusion handling in object detection," in *Multidisciplinary Computational Intelligence Techniques: Applications in Business, Engineering, and Medicine*. IGI Global, 2012, pp. 61–74.

[48] M. Singh, A. Basu, and M. K. Mandal, "Human activity recognition based on silhouette directionality," *IEEE transactions on circuits and systems for video technology*, vol. 18, no. 9, pp. 1280–1292, 2008.

[49] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[50] P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3d video sequences of people," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 362–381, 2010.

[51] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern recognition*, vol. 25, no. 1, pp. 65–77, 1992.

[52] D. Cunado, M. S. Nixon, and J. N. Carter, "Using gait as a biometric, via phase-weighted magnitude spectra," in *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, 1997, pp. 93–102.

[53] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 878–885.

[54] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *2008 IEEE Conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[55] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, pp. 13–es, 2006.

[56] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36.

[57] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[58] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1395– 1402.

[59] O. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "The oxford-iiit pet dataset," 2012.

[60] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4," *International Journal of Computer Vision*, pp. 1–26, 2020.

[61] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[62] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[63] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[64] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[65] J. R. A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," *Retrieved September*, vol. 17, p. 2018, 2018.

[66] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[67] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in´ context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[68] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[70] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[71] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[72] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[73] Google, "Open image dataset v6," https://storage.googleapis.com/ openimages/web/index.html, 2020, [Online; accessed 25-February2020].

[74] N. S. Punn and S. Agarwal, "Inception u-net architecture for semantic segmentation to identify nuclei in microscopy cell images," *ACM Transactions on Multimedia Computing,*

Communications, and Applications (TOMM), vol. 16, no. 1, pp. 1–15, 2020.

[75] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, Ł. Kaiser, N. Kalchbrenner, N. Parmar *et al.*, "Tensor2tensor for neural machine translation," *arXiv preprint arXiv:1803.07416*, 2018.

[76] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[77] A. Sonawane, "Medium : YOLOv3: A Huge Improvement," https: //medium.com/@anand sonawane/yolo3, 2020, [Online; accessed Dec 6, 2019].

[78] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[79] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 748–756.

[80] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–36, 2018.

[81] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.

[82] N. S. Punn and S. Agarwal, "Crowd analysis for congestion control early warning system on foot over bridge," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, 2019, pp. 1–6.

[83] Pias, "Object detection and distance measurement," https://github.com/ paul-pias/Object-Detection-and-Distance-Measurement, 2020, [Online; accessed 01-March-2020].

[84] J. Harvey, Adam. LaPlace. (2019) Megapixels: Origins, ethics, and privacy implications of publicly available face recognition image datasets. [Online]. Available: https://megapixels.cc/