Visual Object Tracking Using Deep Learning

Mehmet Aydin, Mina Shahmoradi, Osman Yilmaz.

Abstract

As one of the effective methods to prevent the spread of the Covid-19 is keeping the distance between people, in this project we will focus on presenting a solution to observe social distancing in public areas with the help of visual objects tracking methods. Creating solutions in this matter using multi object tracking methods will be a fundamental step for the visual social distancing problem. As a result, this study aims to use current available methods of multi object tracking to present a solution to observe social distancing in public areas.

0. Authors' Contributions

Mehmet Aydin was responsible for researching different methodology used in visual object tracking. He also researched in detail about Deep Sort methodology.
Mina Shahmoradi was responsible for finding the application area and researching the field in detail. She was also responsible to implement FairMOT method.
Osman Yilmaz was responsible for searching about FairMOT methodology in detail. He also performed YOLO3 and YOLO4 methods and captured the result.
The Report was written by all group members.

1. Introduction

As COVID-19 emerged in November 2019, it have put the world into a new state of pandemic with regulations that changed everyone's daily life such as quarantine, curfews, and social distancing. Corona virus is still getting many lives daily and as a threat to human lives we need to practise new regulation as individuals. On the other hand, scientist in different fields are trying their best to figure new ways to overcome the current situation one way or another. As well, data scientists have advanced on creating the new models to analyse the data to reduce virus spreading and represent new ways how to act more efficiently.

Human tracking has been used for surveillance and security in different environments and applications, such as sports events catching burglars, measuring number of attendees in big venues and festivals, etc. In the current situation it can also be used to detect number of people in specific area surpassing the social distancing limits. This approach can be used to minimize contacts at hotspots by alarming the system or to evaluate and reorganize the crowded entrances, walkways, etc so they can be reorganized to prevent such matters.

Object tracking is the acquisition of information such as position, speed, or the direction of predetermined or undetermined objects in a sequence of images or a video. Usually, firstly the objects of the frame are recognized by detecting algorithms, then the object tracking algorithm assigns a specific ID to each object present in consequent frames of the video. Designing an effective object tracking model some steps has more influence, such as the choise of the model for both stages of detecting and tracking and how the algorithm weighs between this two stages.

2. Related Work

In this project we encounter multiple objects tracking as oppose to single object tracking. For multiple objects tracking two different approaches are used, firstly methods based on matching objects using their visual representations (e.g., Deep SORT), secondly methods to combine detection and tracking (e.g., JDE, Tracktor, FairMOT).[1][2][5] Methods that combine detection and tracking also use two fundamentally different approaches as either two-stage model where they do detection and re-identification separately, or a single stage model where a single network does both steps concurrently. Both approaches aim to overcome the re-identification, motion prediction and occlusion problems by using different techniques.[2]

2.1. Deep Sort

Deep sort is one of the simplest algorithms used in object tracking. It is used for tracking multiple objects in real-time applications. Like the SORT algorithm, the Deep SORT algorithm uses the Kalman Filter method to predict the location of objects in the next image. Deep SORT and SORT algorithms are separated from each other by the method they use when associating objects. A convolutional neural network (CNN) is designed for object classification to be used in the Deep SORT algorithm. The convolutional neural network is trained until high accuracy is achieved. Thanks to CNN, the most distinctive feature of the object to be classified, and the most distinctive feature that distinguishes the object from other objects, is tried to be determined. In the last layer of the CNN structure, the classification of the objects is done. This classification

process is done according to a vector representing the object. In the Deep SORT algorithm, a vector is obtained by passing each detected object through the neural network and using these vectors to associate the two objects. This vector is called the "appearance feature vector" According to the SORT algorithm, the Deep SORT algorithm is more successful in object association in cases such as occlusions since it is more specific and distinctive features of the object are examined to associate an object with another object.

2.2. FairMOT

FairMOT has been designed based on two homogeneous layers for detecting and reassigning ID to objects detected. This algorithm overcomes some of the obstacles in the way of real time tracking of multiple objects, for example it will treat re-ID task as equal as detection, which will be a better solution comparing it to anchor based algorithms. In Ancher based algorithms although detection of objects might be done perfectly, but as in each frame large number of re-ID tasks should be done (as a secondary priority of the algorithm) it will result in a rather poor tracking result. Using an anchor free style, FairMOT uses a position-aware measurement map to address each object by its center and size, likewise it will assign an ID to the object centered in each pixel in order to characterize it. Taking advantage of this method will result FairMOT to optimize the re-ID assignment as well as detecting used in the multi object tracking in a crowded scene. [5]

2.3. Social Distancing

There are approaches where an augmented circle is drawn around each person to visually see whether the distance is kept or not [4], but our approach will be to use the average personal space in square meters in order not to detect small friend groups or families as violation of social distancing. One of the reasons to decide on this metric is the allowance of group meetings of at most 10 people in Finland. In order to observe the average personal space in square meters and take necessary measures in case the number of violations throughout the day is exceeding certain limit constantly, this project aims to use surveillance cameras of normally crowded public areas.

3. Methodology

As a methodology we have used Deep-SORT. It is one of the simplest algorithms used in object tracking. It is used for tracking multiple objects in real-time applications. In figure 1 we first detect the objects and their possible location using YOLOv4 [8] with pre-trained data. The difference detector masks the frame in order to see the region of interest; this can reduce the computational cost and speed up the procession of detection or tracking. [7] On the other hand, in the tracking side Kalman filter and two association metrics welcome us with respect to Hungarian algorithm.

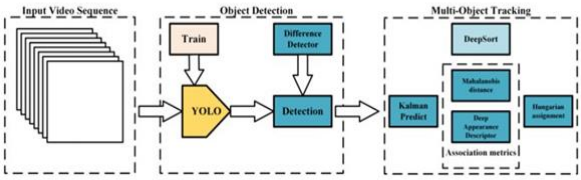

Figure 1: The workflow of real-time detection and tracking [7]

3.1. Object Detection

3.1.1. YOLOv4

One of the most prominent object detection methods nowadays is You Only Look Once (YOLO) which was presented by Redmon et al. [22] in 2016. YOLO is a method which is used to detect or track single or multiple objects that are in the camera's field of view. After coming into the computer vision scene, YOLO immediately got a lot of attention by fellow computer vision researchers. In our project, the latest version of YOLO (YOLOv4) [6] will be modified such that it detects all the necessary objects that appear as ground truth sound sources in the recorded scenes of the project.

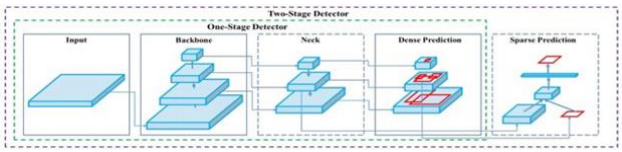

Figure 2: Architecture of object detection of YOLOv4 [8]

3.1.1.1. Input

It is the layer on which the image is given to the model. The input can be an image, patches, or an image pyramid.

3.1.1.2. Backbone

The backbone is the layer where feature extraction is performed. Useful, calculable information is obtained from the information contained in the image. It has been observed that the CSPDarknet53 feature extraction model gives better results in YOLOv4 [8]. The other feature extractor can be VGG16 [9], ResNet50 [10], ResNeXt-101, [21].

3.1.1.3. Neck

It is the layer between the backbone and the head. In order to obtain more information while estimating the objects, an intermediate layer called the neck has been added. Here, detailed information is extracted from neighboring feature maps with information from the bottom-top or top-down flow. In YOLOv4 instead of FPN

Spatial pyramid pooling (SPP) [14], Path aggregation method (PANet) [12], Spatial attention module (SAM) [13] are used.

3.1.1.4. Head

To proceed in head section, bounding boxes and class of each of them will be estimated. Depending on the state-of-the-art detector stage (whether it is a one stage or two stage detector) this section will be divided up to two parts as well. A region proposal network will be used in two stage detectors which firstly will determine region of interest and then use this data to classify object and bounding box regression. Two step detectors normally have more accuracy rates, although they perform slower than single stage detectors. On the other hand, some models like YOLO (You Only Look Once) and SSD (Singe Shot Multi Box Detector) by taking an image as the input will determine class probabilities and bounding box coordinate as a simple regression problem with lower accuracy rate. [15]

3.1.2. The difference Detector

Difference detector masks the frame in order to see the region of interest, this can reduce the computational cost and speed up the procession of detection or tracking.

3.2. Object Tracking

3.2.1. Kalman Filter

As we mentioned above, in order to track the object, first, object detection should be done. We can use any object detection algorithm of your choice here (YOLOv4 has been used). The detected object is enclosed in a geometric shape and a previously unused number is assigned. The tracking of the object is done with this trick. The velocity of the object detected for the first time is also set to zero.

In deep-sort based tracker. The Kalman filter plays essential role. Kalman tracking can be defined on eight-dimensional state space (x, y, c, h, x', y', c', h') which includes the center position of the bounding box c is aspect ratio and h is the height of the input. The other variables are the relative velocities of the variables.

The location of the object in the next image is determined by the Kalman Filter method and the coordinates are updated accordingly. Kalman filters are used to make state estimates of linear systems with a rough state vector Gaussian distribution. It tries to predict the next location of the tracked object by checking previous and next location. Kalman is a structure that is widely used in object tracking due to their simplicity and speed of the filter structures. They are usually applied on noisy data. It allows estimating as close as possible to the truth from noisy and imprecise data. We are trying to make an estimate close to the real state of the system by obtaining data that you cannot directly measure or whose accuracy is not accurate. If the data you obtain contains noise, the Kalman filter will help you estimate the closest true value for you.

3.2.2 Association Metrics

SORT achieves a good performance in terms of sensitivity and accuracy of the tracking, despite the effectiveness of the Kalman filter, it still returns false positive ID and has a low accuracy in tracking through different viewpoints and occlusions and so on. to solve this, there are two things can be applied these are A distance metric for measuring association and an efficient algorithm for correlating data. The [3] authors decided the square Mahalanobis distance would be helpful to get rid of this problem to include uncertainties from the Kalman filter. We could have very good idea of the true connotation if we match this distance. So far, we have an object detector that gives us the detections, Kalman filter tracks detections and provides us the missing track. Hungarian algorithm that solves the missing traces and correlation problem. Despite the effectiveness of the Kalman filter, because of the occlusions, different viewpoints, non-stationary camera it fails in most real-world scenarios. In order improve this the authors of the [3]. Introduced another method that is called Deep appearance metric. So, a classifier is created based on our data set [11], which has been trained until we achieve good accuracy. then the final classification layer is subtracted from this network that achieved enough good accuracy, leaving a dense layer that produces a single feature vector waiting to be classified, this feature vector is known as the Appearance descriptor.

3.3. Challenges

Some challenges encountered in object tracking of this project was:

- The tracked object is unexpectedly out of view.
- The tracked object passes behind another object and is not visible which mostly encountered problem called occlusion.
- Detecting objects after their intersect, reassigning the correct ID to them.
- Detecting objects as themselves even though the object looks different due to its movement or camera movement.
- Object scale in comparison to the video size can cause detection problem, for example in a camera zoom.
- Lighting changes in video can affect perceived image of objects and result inconsistency to detect a same object in different placement in a video.

Indeed, incorporating all the output probabilities in the computation of the final illuminant color estimate is found to be in general more useful than using only the cluster centroid with the highest probability, as this would help to

reduce the likelihood of making high errors in the illuminant computation of other cases in which CNN may fail to predict the correct cluster.

4. Results

The performance of the proposed method is calculated by the metrics of MOT (Multi Object Tracking) challenge. These widely used evaluation metrics are MOTA (Multi Object Tracking Accuracy), MOTP (Multi Object Tracking Precision), IDSW (ID Switch), IDF1, MT (Mostly Tracked), ML (Mostly Lost) and FPS (Frame Per Second). [17] MOTA, as shown in equation 1, is not a simple accuracy calculation but rather a complementary evaluator of the tracker. Tracker system creates bounding boxes to track the object and if the bounding box intersection over union (IoU) is matching at least %50 of the ground truth, it is a True Positive (TP) object tracking. The threshold can be changed depending on the application. If the IoU is less than %50 of ground truth, it is considered as False Positive (FP). If there exist a ground truth bounding box and the created tracker model failed to detect the object, it is counted as False Negative (FN). True negative (TN) cases are eliminated from the calculation in object tracking because they represent every part of the image that model did not predict an object's existence. MOTP as shown in equation 2, is the altogether ratio of bounding box intersection between predicted ones and ground truth for the frame "t". Therefore, it concentrates more to the quality of the tracking. IDSW is the number of times the tracked object's specified ID number has changed wrongly. Deep SORT algorithm when it was introduced in 2017, decreased the ID switches %45 compared to the state-of-the-art system of that day and it was one of the fundamental reasons to be chosen for this project.[3] IDF1 is the F1 score that is calculated by identification precision and identification recall of the tracker's identification capability over the objects. MT summarizes the situation where the tracker tracked the same object %80 of the frames that object occurred. ML on the other hand is the situation where tracker tracked the object less than %20 of the whole frames that object occurred. FPS at last, is an essential metric to evaluate if the tracker can perform in real time or not. It is the number of frames the tracker processes in 1 second. [17]

$$MOTA = 1 - \frac{(FN + FP + IDSW)}{GT} \quad \in (-\infty, 1] \tag{1}$$

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \tag{2}$$

The performance of the proposed network is compared against other state of the art multi object trackers under MOTA metric. The decision is to use state-of-the-art object detector YoloV4 with Deep SORT in order to successfully track objects. Results showed us that FairMOT performs much better than any of the publicly available trackers when it comes to tracking people in crowded scenes but there were problems with the necessary libraries to run FairMOT. Result of state-of-the-art trackers are shared in table 1, with our own results included as Yolov4 + DeepSORT.

Table 1: Performance comparison between various methods with the specified dataset.[3][5][8][19]

Trackers	Evaluation Dataset	MOTA Score
YoloV4	MOT17	56.46
YoloV4 + DeepSORT	MOT17-04	61
Faster RCNN + DeepSORT	MOT16	61.4
FairMOT	MOT16	74.9
FairMOT	MOT17	73.7

In addition to the metric, the actual results are compared with FairMOT since it is evaluated on the same MOT dataset. [20] Figure 3 and 4 illustrates the difference between our result and FairMOT for the same scenes. When calculating the safe square per meter area per person, it is considered that 1 meter as the radius of the circle is safe, so that 3.14 square meter per person is required. When processing recordings from different security cameras, the covered area of the security camera is provided to the command prompt.

Figure 3: A frame from MOT17-04 dataset with YoloV4 + DeepSORT and FairMOT results [5]

It is understood that the main difference between the models come from the training dataset and anchor box approach. For FairMOT, since it tracks objects with a heatmap and anchor free approach without bounding box, by the help of training it on CrowdHuman dataset, it performs better on crowded scenes.[5] [18] But due to FairMOT being a recent method, it has less community support and has problem with scalability to different operating systems. As for our approach YoloV4 with Deep SORT, it performs better with low quality videos with providing less object ID switches due to MARS dataset trained for deep appearance descriptor. [3] [11]

Figure 4: A frame from MOT20-01 dataset with YoloV4 + DeepSORT and FairMOT results [5]

Discussion Even though FairMOT seems better with the provided results, due to FairMOT being a freshly new method, it has less community support and has scalability problem with different operating systems due to libraries used in the system. On the other hand, one of the main advantages of using Deep SORT is that it can be used with any new state of the art object detectors whenever there is a better option. Therefore, Deep SORT's scalability and community support are much better than FairMOT.

5. Conclusions

In this paper, we have presented a new approach to deal with the spread of the Covid-19 in public spaces and whether the social distancing is occurring or not. For using state-of-the-art object trackers, we decided that implementing personalized area for each person will not be suitable since the togetherness of families and small friend groups are unavoidable. The rules of at most 6 people can be together at the same time of Pirkanmaa region as of March 2020 supports our approach of social distancing by calculating the square meter area per person. We researched state of the art object trackers and their methodologies and decided to implement two methods: YoloV4 with Deep SORT and FairMOT. Due to library errors of "cython_bbox" in Windows operating system, FairMOT failed and tracking by detection methodology of YoloV4 with Deep SORT was implemented. Input of the model was security camera covers of crowded public areas which would count the number of people in each frame in order to calculate the square meter area per person. We have learned about object tracking methods in detail and commonly used datasets to train these tracking networks. We understood that for our purpose, training the tracker with CrowdHuman dataset makes a huge difference due to its similarity with the real scenario of public places. To improve our system even more, FairMOT system can be tried in a different operating system with good GPU infrastructure and system can give an outcome of heatmap of people's location when the social distancing is not kept, and system is alarming.

References

[1] Wang Z., Zheng L., Liu Y., Li Y. and Wang S, 2020, "Towards Real-Time Multi-Object Tracking", arXiv:2005.04813v1.

[2] Bergmann P., Meinhardt T. and Leal-Taixe L., 2019, "Tracking without bells and whistles".

[3] Nicolai Wojke, Alex Bewley, Dietrich Paulus : Simple Online and Realtime Tracking with a Deep Association Metric

[4] Cristani M., Bue A.D., Murino V., Setti F. and Vinciarelli A., 2020, "The Visual Social Distancing Problem".

[5] Zhang Y., Wang C., Wang X., Zeng W. and Liu W. "FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking", 2020

[6] Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. 2020. arXiv: 2004.10934.

[7] Xu Zhang, Xiangyang Hao, Songlin Liu, Junqiang Wang, Jiwei Xu, Jun Hu, "Multi-target tracking of surveillance video with differential YOLO and DeepSort," Proc. SPIE 11179, Eleventh International Conference on Digital Image Processing (ICDIP 2019), 111792L (14 August 2019); doi: 10.1117/12.2540269

[8] Alexey Bochkovskiy,Chien-Yao Wang,Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection ",2020

[9] K. Simonyan, A. Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition

[11] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang and Qi Tian: MARS: A Video Benchmark for Large-Scale Person Re-Identification

[12] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, Jiaya Jia: Path Aggregation Network for Instance Segment. 18 september 2018
[13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon : CBAM: Convolutional Block Attention Module.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg: SSD (Singe Shot MultiBox Detector)
[16] Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, Mubarak Shah: Multi-Target Tracking in Multiple Non-Overlapping Cameras using Constrained Dominant Sets
[17] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, Francisco Herrera : Deep Learning in Multi-Object Tracking: A Survey
[18] Shao, Shuai and Zhao, Zijian and Li, Boxun and Xiao, Tete and Yu, Gang and Zhang, Xiangyu and Sun, Jian: CrowdHuman: A Benchmark for Detecting Human in a Crowd
[19] Omar Moured: Evaluation of Deep Learning Based Multi Object Trackers
[20] Multi Object Tracking Benchmark
[21] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He: Aggregated Residual Transformations for Deep Neural Networks
[22] Redmon, J., Divvala, S., Girshick, R. B. and Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), pp. 779–788.