CROWD DETECTION AND TRACKING FROM MULTI-CAMERA SURVEILLANCE VIDEOS

1st Sarika Sasi.A
M.Tech, Artificial Intelligence and Data Science
Muthoot Institute Of Technology And Science
Kochi, India
sarika.sivadam0@gmail.com

***Abstract*—The importance of video-based monitoring systems is growing, causing computer vision to become more and more popular. With the increase in human population, it is necessary to monitor the crowd, whether in a public or an industrial setting. The main aim of the project is to analyse the crowd from multiple surveillance CCTV video. Crowd analysis opens up a new application domain, such as the automatic detection of chaotic acts in crowds and the location of anomalous zones in pictures. Monitoring people's behaviour is highly beneficial due to the ubiquitous presence of multiple cameras in surveillance systems. Multiple perspectives of the scene enable for the handling of occlusions and sensor failures. There are three stages in this project detection, tracking and matching the tracked person. The initial stage in developing intelligent surveillance applications is to monitor objects using multiple cameras. The YOLO algorithm is used to detect and recognise objects. For tracking, the DeepSORT method is utilised, which is simply a Deep association metric combined with the SORT algorithm. Histogram A framework for crowd detection, zonal counting, tracking of people and behaviour analysis of crowd from multi-camera surveillance video is presented in this model.**

***Index Terms*—multi-camera, surveillance, crowd, detection, tracking**

I. Introduction

A crowd is a group of people or something that is part of a community or society. The crowd phenomenon is well-known in many fields of study, including sociology, civil engineering, and physics. The world's population is expected to rise at a rate of 1.05 percent per year by 2020 [10] (down from 1.08 percent in 2019, 1.10 percent in 2018, and 1.12 percent in 2017). The current annual population growth rate is expected to be 81 million people. Crowd analysis opens up a new application domain, such as the automatic detection of chaotic acts in crowds and the location of anomalous zones in pictures. Multiple perspectives of the scene enable for the handling of occlusions and sensor failures. The initial stage in developing intelligent surveillance applications is to monitor objects using multiple cameras. This research topic tries to better understand how people behave in huge groups and to extract useful information from recordings with enormous crowds of people.

One of the fundamental areas in surveillance computer vision research is crowd analysis. Because it offers a wide range of applications in video surveillance, crowd analysis and scene interpretation has gotten a lot of interest recently. With

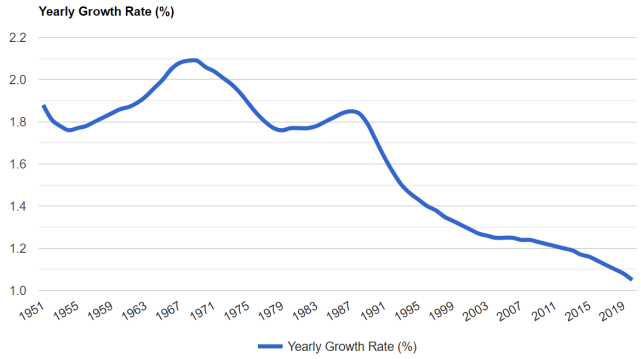

Fig. 1. Population Growth Rate

the expanding population the necessity of crowd monitoring is high. The local authorities have to monitor people, but manual surveillance is tedious and time consuming process. Crowd analysis includes detection,counting and analysis of movement. Here the objective is to design a model for crowd detection and behavior analysis of the crowd from multi-camera surveillance visuals. Multi-camera surveillance camera video covers larger area than single camera videos. Multi-camera tracking systems automatically detect and track people through a network of cameras. This paper proposes an approach to detect and track the people present in multi-camera scenarios.

II. Literature survey

The survey conducted by Shareef et.al. [1] aims to provide a YOLOv4-based model for monitoring social distancing when the Corona virus started spreading over the world. Social distancing (SD) is one of the most efficient ways to stop COVID-19 from spreading since it encourages people to keep their distance from one another. The model starts with a video or image as input and generates Social Distancing violation notifications. Based on deep learning techniques, the YOLOv4 is utilised in this model to detect people in public settings such as streets, malls, railway stations, and universities. When a violation occurs, the model employs a specified Social Distancing Threshold (SDTH) and a violation index (VI) to determine when it occurs and then triggers a

warning sub-system to make an immediate awareness. The authors also conducted a thorough investigation and discussion of the existing SD literature, object detection methods, and SD monitoring.

Ullah, Habib et.al [2] adopted a two-stream convolutional architecture that integrates spatial and temporal networks, offering a unique method for modelling crowd video dynamics. It creates a unified deep model for crowd video modelling that incorporates both geographical and temporal information. By modelling geographic information efficiently and recording temporal information compactly, the model lowers the complexity of spatial and temporal fluctuations. It is unaffected by the assumption of constant crowd flow. The model's spatial component works on individual frames, effectively collecting crowd video modelling features from them. Through dense optical flow, a motion flow field is extracted from the movie.

Multi-target multi-camera tracking (MTMCT) model is proposed [5] by Xu, Jian et.al. which aims to identify and track every person appearing in videos. Multi-target multi-camera tracking systems track a large number of persons in multiple camera footage. Detection, feature extraction, and data association are the three steps. It also includes minor post-processing techniques like pruning and interpolation. Authors focused on the data association process. The feature group approach was introduced to offset the loss of accuracy due to occlusions. Feature grouping is a technique introduced for improving the speed and accuracy of data association. To begin, describe the flaws of the baseline data association approach. These pedestrians are given different IDs, and each ID serves as a cluster's unique label.

In the study conducted by Feng et.al. discusses the flow of multiple object tracking based on detection and their abnormality [6]. There is a detector and a tracker. The detector takes a frame from video stream and detect the object and then detection result is send to the tracker to get the multiple object trajectory. Yolov3 is adopted as the detector to do the detection part. Giving an input image, it can output the locations of interest, which is the pedestrian in this method. It works using a trained convolution neural network (CNN) and Deep-sort algorithm is used as the tracker. The locations of the pedestrians is given by the detector as input. It uses a Kalman filter with constant velocity motion and linear observation model to calculate the location of every pedestrian in the video. Then pre-trained CNN is applied to compute bounding box in order to find the most possible location of the pedestrian. Then it associates the position of pedestrians in the previous and subsequent frames for tracking. Multiple object tracking gives the location and motion trajectory of every person in the video, which will be used to do abnormality [7] detection and analysis.

For monitoring the social distancing in sustainable smart cities using deep learning a data-driven deep learning-based framework is introduced by Shorfuzzaman et.al. [4]. Use of real-time object de- tection models such as YOLO, SSD, and Faster R-CNN[6]. Faster R-CNN was built from two of its predecessor architectures, called R-CNN and Fast R-CNN where Region of Interest(ROI) are generated using selective search algorithm. Researchers introduced the SSD architecture to perform object detection by combining the region proposal and feature extraction in a single deep neural network. YOLO is one of the fastest object detection algorithms capture videos. Perspective transformation of real-time video to transform into bird's eye view [4]. This bird's eye view is computed by a uniform distribution of points in both horizontal and vertical directions, even though the scale is different in each direction.

III. METHODOLOGY

Crowd analysis deals with a broad range of topics, including crowd detection, crowd tracking, crowd counting, pedestrian traveling time estimation, crowd movement analysis and crowd behavior analysis. In this project, multi-camera crowd detection and tracking of pedestrians using YOLOv4 and DeepSORT are discussed. This method consist of four phases, i.e., detecting the object, tracking the object ,counting the object in the zone and movement analysis.

Fig. 2. Work Flow

The input video sequence is taken from multiple cameras. Where there is a primary camera and secondary camera. The dataset is taken from "EPFL" data set. Detection of pedestrians from the video using YOLOV4 algorithm. This will give the output with bounding box. Tracks the pedestrians by giving unique ID and Deep SORT algorithm is used for that. DeepSORT is applied to primary camera and then Zonal counting to count the number of people. Movement trajectory of people present in the video is analyzed from that behavior of people in that is analyzed. Detection and tracking and tracking of people in the videos play an important role in this model. The model is able to alert the system when the number of people present in a particular area is greater than a threshold. The sudden change in the movement of the people in that area can be analyzed and compute the abnormality if it is greater than a threshold. The objects in the box's frame are detected and tracking is done. An unique ID is created for each of the detection. Then, track the object over a period of time till it remains in the frame. The number of people present in each zone is counted. After that, a matching sequence is set to be performed for similar detection.

The input video sequence will be made in to frames then YOLOV4 algorithm is applied to both the videos from that the person are identified and a bounding box will be given to each person in the video. The image of the person detected with bounding box will be cropped. Deep SORT algorithm needs well detected images, the cropped image which is detected by YOLOV4 algorithm will be given as input. Deep SORT is applied to primary video, then it matches with the secondary video cropped images. For matching the IDs histogram is used.

Given ID for each person and match person present in different frames by using its features. These images of the person with their unique is saved. Tries to match the person by analyzing its feature here histogram is measured. Then as a result the matched person will have same ID. Deep SORT also gives trajectory motion.

A. Detection

The detection of objects is the focus of this phase. YOLOv4 is used in conjunction with pre-trained weights to detect objects. Unlike sliding window or region-based approaches, YOLO sees the full image during training and testing, so it implicitly stores contextual information about classes as well as their appearances. To forecast each bounding box, YOLO takes characteristics from the entire image. The model's inference time has increased significantly in YOLOv4. The head's major task in YOLOv4 is prediction, which includes categorization and regression of bounding boxes. YOLOv4 is a more advanced version of YOLOv3, with increased speed and accuracy. All of the YOLO models are created with the goal of detecting objects. To classify the objects in an image, object detection models are trained. When object classes are identified, they are contained within bounding boxes and are categorised. The COCO (Common Objects in Context) dataset, which contains a wide range of 80 object classes, is commonly used to train and evaluate object detection models. The following are the primary arguments for utilising YOLOv4 for detection: For real-time object detection, YOLOv4 prioritises it.

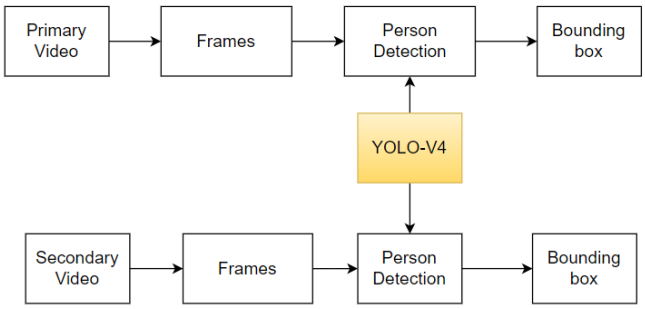

Fig. 3. YOLOv4 Object Detection

B. Tracking

Deep SORT algorithm is used for tracking. It is an advanced version of the Simple Real-Time Tracker (SORT) algorithm. DeepSORT track not just using distance, it also uses velocity, what that person looks like. It enables for the addition of features by computing deep features for each bounding box around an object and factoring deep feature similarity into the tracking algorithm. This algorithm is applied to primary video and is not applied to secondary camera video. The people in secondary video is matched by using the IDs Every time a new frame is loaded, the position of each track is calculated based on its prior positions. For track estimation, just the spatial information is needed. The appearance feature vector is then used to characterise all of the features of a given image. By extracting features and tracking images from previous frames, this vector acquires the appearance information of detection. It extracts features in such a way that features with various identities are separated by a significant amount of space, while features with the same identity are near together. Using the appearance feature vector and the tracks' expected position, the new detection results are connected with each approaching frame's existing tracks.

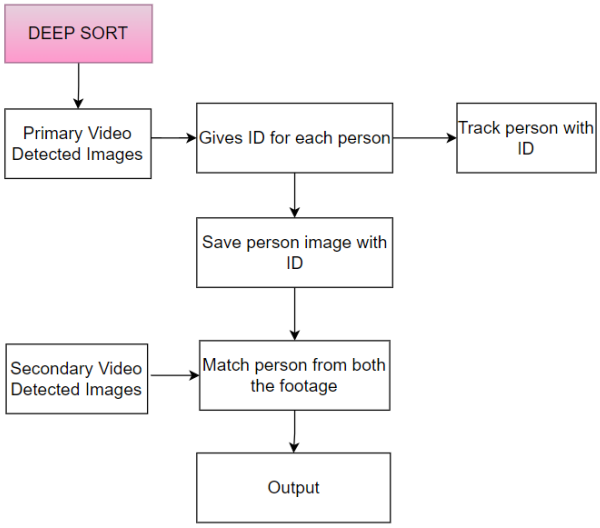

Fig. 4. Tracking using DeepSORT

C. Counting

For counting the number of people in the video zonal counting method is used. Setting the coordinates in the frame creates a zone. There will be four zones in the frame. To calculate the number of objects in the zone, count the number of objects with the unique ID. When tracking finds a specific object, it assigns that thing a unique id number, which then becomes the only source for tracking that specific object. It is easier for the tracker to maintain track of objects throughout the video or image series by issuing unique IDs.

IV. Result

Multi-camera videos are taken from "EPFL" data set, two camera videos are selected from a scenario and one video is considered as primary video and other secondary video. For object detection YOLOv4 is used, bounding box is given to each person in the video. After applying DeepSORT to primary video each person has unique id with different colored bounding box. The frame is divided in to four zones the number of people in each zone is displayed in white color if the number of people in a zone is greater than a threshold it will be in red color. The movement of people in the video is displayed and if it is greater than the threshold a warning will appear. People in primary video is matched with secondary

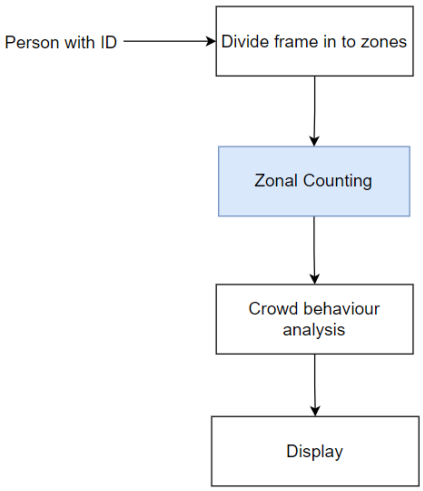

Fig. 5. Zonal counting

video using histogram, it is a form of color matching method. This sometimes leads to mismatching if two people has same color dress or any other features.

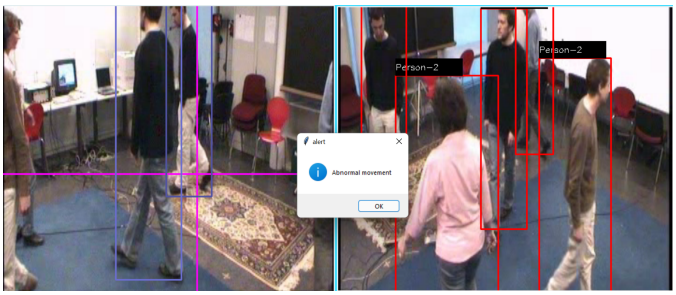

Fig. 6. Abnormality alert

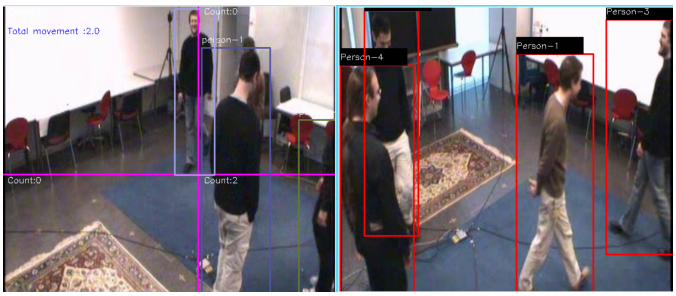

Fig. 7. Detection and Tracking

V. Conclusion and Future Scope

Multi-camera crowd analysis has different phase detection, tracking, people counting and movement analysis. The people detected in the primary is matched using histogram with the second video defines the performance of the system. Limitation of this system is that it leads to mismatching of IDs. This system can be used for future references and work on the matching of people in the videos. The abnormal behaviour of the whole area is displayed in this project in the future the work to analyze the behaviour of each person in the video.

References

[1] Shareef, Ahmed Abdullah A., et al. "YOLOv4-Based Monitoring Model for COVID-19 Social Distancing Control." Smart Systems: Innovations in Computing. Springer, Singapore, 2022. 333-346.
[2] Ullah, Habib, et al. "Multi-feature-based crowd video modeling for visual event detection." Multimedia Systems 27.4 (2021): 589-597.
[3] Xu, Jian, Chunjuan Bo, and Dong Wang. "A novel multi-target multi-camera tracking approach based on feature grouping." Computers Electrical Engineering 92 (2021): 107153.
[4] Shorfuzzaman, Mohammad, M. Shamim Hossain, and Mohammed F. Alhamid. "Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic." Sustainable cities and society 64 (2021): 102582.
[5] Xu, Jian, Chunjuan Bo, and Dong Wang. "A novel multi-target multi-camera tracking approach based on feature grouping." Computers Electrical Engineering 92 (2021): 107153.
[6] Feng, Fujian, et al. "Abnormal Crowd Behavior Detection Based on Movement Trajectory." Chinese Conference on Image and Graphics Technologies. Springer, Singapore, 2020.
[7] Ciaparrone, Gioele, et al. "Deep learning in video multi-object tracking: A survey." Neurocomputing 381 (2020): 61-88.
[8] Xie, Shaoci, Xiaohong Zhang, and Jing Cai. "Video crowd detection and abnormal behavior model detection based on machine learning method." Neural Computing and Applications 31.1 (2019): 175-184.
[9] Amir Sjarif, Nilam Nur, et al. "Crowd analysis and its applications." International Conference on Software Engineering and Computer Systems. Springer, Berlin, Heidelberg, 2011.
[10] https://www.un.org/en/global-issues/population
[11] https://www.epfl.ch/labs/cvlab/data/data-pom-index-php/