A Novel Social Distancing Analysis in Urban Public Space: A New Online Spatio-Temporal Trajectory Approach

Jie Su[a], Xiaohai He[a, *], Linbo Qing[a], Tong Niu[a], Yongqiang Cheng[b], Yonghong Peng[c, *]

[a] *College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan, 610064, China*

[b] *Department of Computer Science and Technology, University of Hull, Hull, HU6 7RX, United Kingdom*

[c] *Department of Computing and Mathematics , Manchester Metropolitan University, Manchester, United Kingdom*

**Joint Corresponding Authors (Xiaohai He: hxh@scu.edu.cn, Yonghong Peng: Y.Peng@mmu.ac.uk)*

Abstract

Social distancing in public spaces plays a crucial role in controlling or slowing down the spread of coronavirus during the COVID-19 pandemic. The Visual Social Distancing (VSD) offers an opportunity for real-time measuring and analysing the physical distance between pedestrians using surveillance videos in public spaces. It can provide evidence for implementing effective prevention measures of the epidemic. The existing VSD methods developed in the literature are primarily based on frame-by-frame pedestrian detection, which addresses the VSD problem from a static and local perspective. In this paper, we propose a new online multi-pedestrian tracking approach for spatio-temporal trajectory and its application to multi-scale social distancing measuring and analysis. Firstly, an online multi-pedestrian tracking method is proposed to obtain the trajectories of pedestrians in public spaces, based on hierarchical data association. Then, a new VSD method based on sptatio-temporal trajectories is proposed. The proposed method not only considers the Euclidean distance between tracking objects frame by frame but also takes into account the discrete Fréchet distance between trajectories, hence forms a comprehensive solution from both static and dynamic, local and holistic perspectives. We evaluated the performance of the proposed tracking method using the public dataset MOT16 benchmark. We also collected our own pedestrian dataset "SCU-VSD" and designed a multi-scale VSD analysis scheme for benchmarking the performance of the social distancing monitoring in the crowd. Experiments have demonstrated that the proposed method achieved outstanding performance on the analysis of social distancing.

Keywords: Visual Social Distancing; Hierarchical Data Association; Multi-Pedestrian Tracking; Spatio-Temporal Trajectory; Discrete Fréchet Distance; Crowd Gathering

1. Introduction

The COVID-19 pandemic is ravaging the world, which has sadly caused a significant loss to human life, and a great negative impact on society and the economy. On 30 January 2020, the World Health Organization (WHO) declared that the outbreak of COVID-19 constitutes a Public Health Emergency of International Concern (PHEIC) [1]. On 6 January 2021, it reported that there have been 84,780,171 confirmed cases of COVID-19, including over 1,853,525 deaths globally [2]. The rapid spread of COVID-19 is mainly through close contact from people to people, and asymptomatic carriers can also spread the virus to others [3]. Due to the high density and mobility of the urban population and the complexity of the urban environment, the spread of the pandemic has been exacerbated to some extent, which brings to severe challenges to construction, governance and sustainable development of cities.

For the epidemic diseases, measures are taken to prevent and control infections include vaccination, treatment, quarantine, isolation, and prophylaxis [4]. However, the vaccine for COVID-19 has not yet entered the promotion stage, and the more contagious coronavirus variant has been detected. In this scenario, one effective way to control or slow down the spread of coronavirus is to make sure people maintain social distancing in public places. There exists some work in the literature studying the impact of social distancing on the progression of the coronavirus [4-7][8][9]. Using Wuhan as a case study, Prem et al. [5] stated that physical distancing based non-pharmacological interventions have a high potential for flattening the peak of COVID-19 and reducing the overall number of cases. Cacciapaglia et al. [7] demonstrated that social distancing measures are more efficient than border control in delaying the epidemic peak. Sun et al. [8] researched the efficacy of social distancing and ventilation effectiveness in preventing COVID-19 transmission. With the current epidemic unlikely to end in the short term, keeping a safe social distancing[1] from others in public spaces and workplaces is one of the key measures for maintaining a low risk of infection.

In recent years, with the deepening of the concept of "smart sustainable cities" [10][11], countries around the world have deeply integrated information technologies with the various needs of urban development. In this context, some research work has explored ways to prevent and respond to the ongoing COVID-19 pandemic by using urban

[1] The World Health Organisation advises maintaining at least 1 metre (3 feet) distance between yourself and others (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public). In this paper, the safe distance threshold is set to 2 metres (6 feet).

infrastructures and emerging technologies [12][13][14][15][16][17], especially in the aspect of automatic social distancing monitoring in public places [17][18][19] [20][21][22][23][24]. This helps to enhance the resilience and sustainability of cities. The construction of smart cities has also resulted in an explosive growth in video data taken from public spaces. Compared with other big data utilised in existing researches, the video data contains wealthy spatial and temporal information about human. Exploiting video data to study and analyse human trajectories can more precisely mine human activities in various complex scenes, which is an excellent supplement to non-visual big data and has unique advantages. Therefore, during the pandemic, it is of great theoretical significance and research value to measure and analyse the social distancing between pedestrians based on their spatio-temporal trajectories using surveillance videos in public places and take appropriate epidemic prevention measures according to the crowd gathering situations. This research topic is called Visual Social Distancing (VSD), which refers to approaches relying on video cameras and other imaging sensors to analyse the proxemic behaviour of people [18].

In this paper, a new VSD method based on the human spatio-temporal trajectory has been proposed to quantify and analyse the social distancing between pedestrians in public spaces. The contributions of our work are summarised as follows:

1) A new hierarchical association based online and real-time multi-pedestrian tracking method is proposed to obtain pedestrians' trajectories, which can effectively reduce the number of identity switches while achieving overall competitive performance.
2) A new VSD method based on spatio-temporal trajectories is proposed considering both the Euclidean distance between sampling points of trajectories from a local perspective and the Fréchet distance between trajectories from a holistic view.
3) A multi-scale social distancing analysis scheme is proposed, including four evaluation metrics, which can evaluate the crowd gathering situation from various time scales respectively.

The rest of the paper is organised as follows. Section 2 gives the related work of VSD and multiple-object tracking, xxxxx

2. Related Work

2.1 VSD Problem

Some new research work has been conducted to study the VSD problom for COVID-19 [18][19][17][20][21][22][23][24]. For instance, Cristani et al. [18] proposed a VSD method based on body pose estimation. In each frame, the body pose detector is used to detect visible people. In the corresponding bird's eye view (top view), each detected pedestrian is regarded as the centre of the circle, and the safe distance as the radius. Then the VSD issue is converted to a sphere collision problem. Yang et al. [20] proposed a VSD and critical social density detection system to avoid overcrowding by modulating inflow to the region-of-intrest (ROI). Shorfuzzaman et al. [21] used deep learning-based object detection models to detect individuals and implement social distancing monitoring. The Landing AI Company[2] developed a social distancing detection tool by detecting pedestrians in real-time video streams and measuring the distance between pedestrians in the corresponding bird's eye view frames. These methods have made some useful contributions to VSD in the pandemic, but most of them are based on frame-by-frame pedestrian detection rather than pedestrian tracking over a period of time. Although there existed some VSD studies leveraging both detection and tracking approaches [23][24], the tracking algorithms in these methods were employed for tracking already identified people and assign IDs rather than for trajectory-based social distancing measuring. As a consequence, these frame-by-frame distance metric based VSD methods fall in the category of detection-based VSD, while the proposed VSD method based on spatio-temporal trajectories distance metric is a trajectory-based VSD. To the best of our knowledge, this research work is the first attempt to address the VSD issue in a dynamic and spatio-temporal manner.

The distinction between the detection-based VSD and the trajectory-based VSD is shown in Figure 1. The detection-based VSD detects and calibrates the positions of pedestrians, and measures the distance between them frame-by-frame in the bird's eye view. The trajectory-based VSD tracks pedestrians and calibrates trajectories, and metrics the distance between corresponding calibrated trajectories in the three-dimensional spatio-temporal coordinate (adding a time axis *t*). The detection-based VSD method is from a static and local perspective, while the trajectory-based VSD method is from a dynamic and spatio-temporal perspective. However, during the pandemic, the issue that should be considered is the continuous measurement and analysis of social distancing rather than a specific moment. Therefore, it is more

[2] https://landing.ai/landing-ai-creates-an-ai-tool-to-help-customers-monitor-social-distancing-in-the-workplace/

sensible to investigate the VSD problem based on the spatio-temporal trajectory over a time period.

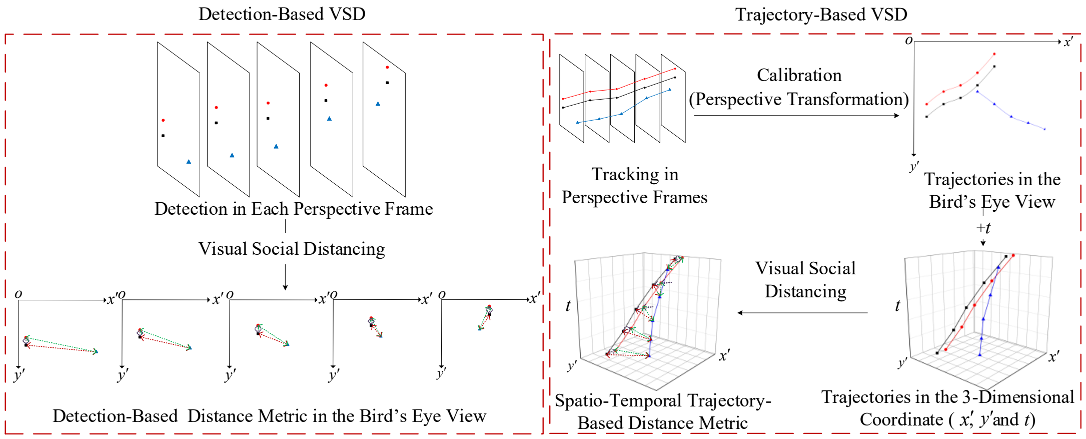

Detection-Based VSD
Trajectory-Based VSD
Detection in Each Perspective Frame
Visual Social Distancing
o
x'
y'
Detection-Based Distance Metric in the Bird's Eye View
Calibration
(Perspective Transformation)
Tracking in
Perspective Frames
Trajectories in the
Bird's Eye View
+t
Visual Social
Distancing
t
Spatio-Temporal Trajectory-
Based Distance Metric
Trajectories in the 3-Dimensional
Coordinate (x', y'and t)

Figure 1 The Difference between the Detection-Based VSD and the Trajectory-Based VSD

2.2 Multi-Object Tracking

The Multi-Object Tracking (MOT) is a fundamental research topic in the field of computer vision, which is widely applied to smart surveillance, autonomous driving, security and other areas. MOT is also an underpinning technique for trajectory-based VSD. In recent years, with the dramatical improvement of detectors [25][26][27], Tracking-by-Detection [28][29][30][31][32][33] has become the mainstream paradigm of the MOT. For Tracking-by-Detection, objects are detected and localised in each frame firstly, and then tracking is conducted by using data association to link detections into trajectories. Therefore, the tracking performance is highly dependent on the performance of the detector and the data association method. Also, MOT can be divided into online tracking [28][29][30][31][32][33] and offline tracking [34] [35][36][37]. Online tracking refers to data association based only on the past and the current frames, while offline tracking refers to data processing by exploiting all the frames or batch frames. Because offline tracking methods demand the entire set of videos to be obtained in advance, they are less favored by real-time tasks compared to their online counterparts.

With regard to online Tracking-by-Detection, the traditional methods include Multiple Hypothesis Tracking (MHT) [38], the Joint Probabilistic Data Association Filter (JPFAF) [39]. Recently, deep learning based methods enchance the tracking performance in complex scences drastically. The Person of Interest (POI) [40] method introducd the high-performance detection and deep learning based appearance feature into the context of MOT. Depending on Convolutional Neural Network (CNN) based detection, the Simple Online and Realtime Tracking (SORT) [29] method utilised the Kalman filter for frame-by-frame prediction and the Hungarian method for data association, by calculating intersection-over-union (IOU) distance as the assignment cost. The Deep SORT [32] method further introduced deep appearance features and motion features on the basis of SORT [29] for assigenmet costs calculation. Also, some research work in the literature focuses on researching hierarchical data association to improve the reliability of association. Bae et al. [30] proposed a hierarchical association method based on the tracklet confidence, which built optimal tracklets by sequentially linking tracklets and detections using the high and low confidence association. Alshakarji et al. [33] proposed a three-step cascade scheme for efficient data association.

3. The Proposed Approach

The proposed approach consists of two main steps: (1) a hierarchical association based online multi-pedestrian tracking method to obtain trajectories of pedestrians; (2) a trajectory-based social distancing measurement and analysis method to evaluate social distancing situations between pedestrians in public spaces.

3.1 Hierarchical Association Based Online Multi-Pedestrian Tracking

Considering the real-time requirement of the trajectory-based VSD task, it is necessary to design a simple and highly real-time tracking method. Inspired by SORT [29] and Deep SORT [32] method, we utilise the Kalman filter and the Hungarian algorithm [42] to address the MOT task. Besides, we design a new hierarchical data association scheme to ensure tracking performance and fewer ID switches.

3.1.1 The States of the Tracklets and the Transition Mechanism

For online Tracking-by-Detection, the essence is a frame-by-frame data association based on detection responses pre-generated by the detector. The detection responses in the current frame are assigned to the existing tracklets (the tracklet is a part of the trajectory formed during the tracking process) according to the data association method. However, the issues of misdetection, occlusion, the appearance and the disappearance of tracking objects lead to many challenges of the MOT task. To tackle these challenges, we adopt a hierarchical data association method based on the states of the tracklets to address multi-pedestrian tracking. According to the number of consecutive associated frames, the states of the tracklets are classified into four categories, namely initial, tentative, stable and deleted. The transition mechanism of the four states of the tracklets is shown in Figure 2.

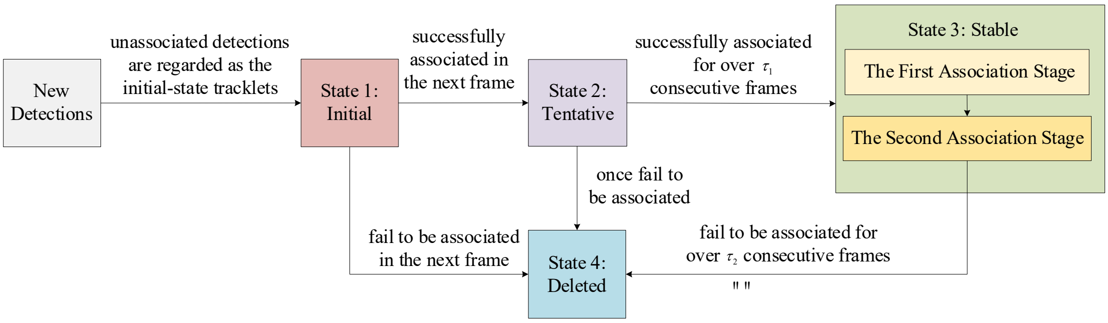

New Detections
unassociated detections are regarded as the initial-state tracklets
State 1: Initial
successfully associated in the next frame
State 2: Tentative
successfully associated for over τ_1 consecutive frames
State 3: Stable
The First Association Stage
The Second Association Stage
once fail to be associated
fail to be associated in the next frame
State 4: Deleted
fail to be associated for over τ_2 consecutive frames

Figure 2 The Transition Mechanism of the Four States of the Tracklets

① The Initial-State

The initial-state is defined for a new detection that cannot be associated with any existing tracklet. In this state, it is regarded as a new tracklet. If it is associated successfully in the next frame, its state will become as tentative. Otherwise, it will be deleted.

② The Tentative-State

The tentative-state is a state when the tracklet in initial-state is successfully associated in the next frame. In this scenario, if the tracklet in tentative-state continues to be successfully associated for over τ_1 consecutive frames (the threshold τ_1 is a small positive integer), its state will progress to stable. Otherwise, it will be deleted.

③ The Stable-State

The stable-state is defined as a state when the tracklet in tentative-state is successfully associated for over τ_1 consecutive frames. Once the state of the tracklet becomes stable, it can only be deemed to be finished and deleted when it fails to be associated for over τ_2 consecutive frames (The threshold τ_2 is a positive integer significantly greater than τ_1). Besides, the stable-state can be further divided into two stages, the first association stage and the second association stage. These two stages employ different feature metrics for assignment cost calculation during the association (Section 3.1.3).

④ The Deleted-State

Under the following conditions, the state of the tracklet will be defined as deleted-state:

(1) when the tracklet is in initial-state or tentative-state, and it fails to be associated in the next frame; (2) when the stable-state tracklet fails to be associated for over τ_2 consecutive frames.

The setting of the initial-state and the tentative-state can effectively address the misdetection issue. This is because the tracklet of the misdetected object in an initial or tentative state will be deleted once the association fails. The stable-state setting takes into account the impact of occlusion during tracking. The tracklet may fail to be associated during the tracking process due to occlusions. Under this situation, the object could not be considered to have disappeared, nor should the tracklet be deleted immediately. The setting of the threshold τ_2 is beneficial to improve the continuity completeness of tracking. And the two association stages of the stable-state helps to improve the reliability of data association. The deleted-state setting is to reduce unnecessary calculations.

3.1.2 The AAM-Softmax Appearance Feature Descriptor

The distance between appearance features is taken into account in the analysis of data association. Inspired by the Cosine Softmax methods [32][43] and the Additive Angular Margin Softmax (AAM-Softmax) methods [44][45], an AAM-Softmax appearance feature descriptor is designed to obtain well-discriminative appearance features of pedestrians. Before being applied in the online tracking task, the descriptor is trained offline by using a large-scale person re-ID dataset Market1501 [46] with 12,936 training images of 751 identities, which facilitates deep metric learning in a pedestrian tracking context.

We mainly use convolutional layers and residual blocks [47] to consturct the architecture of the proposed descriptor network network (shown in Figure 3). The pedestrian images with size 128×64×3 are input into the CNN based architecture, including two convolutional layers (each layer has 32 kernels with size 3×3 and stride 1), a max-pooling layer (pooling size is 3×3 and stride is 2) and six residual blocks with two stacked layers. Through the CNN architecture, the feature maps with size 16×8×128 can be obtained. Then after a Global Average Pooling (GAP) layer, a Batch Normalisation (BN) layer and a l_2 Normalisation layer, the descriptor finally outputs a feature vector with 128 dimensions. In the training phase, the ID of each sample is utilised as a training label, and each embedded feature is input into the fully connected layer followed by the AAM-Softmax classifier, performing supervised learning by the AAM-Softmax loss.

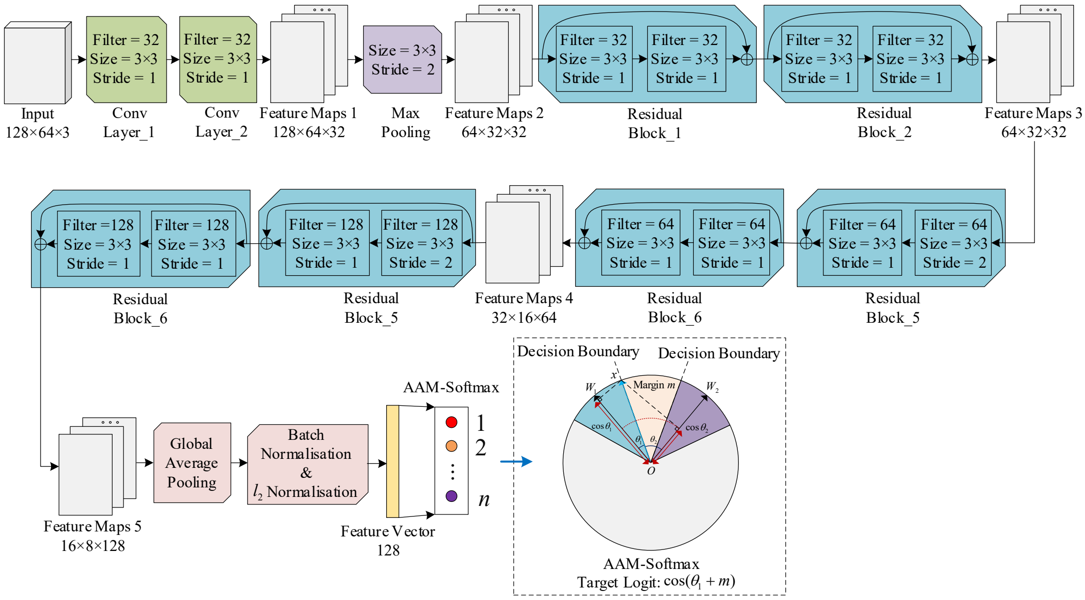


Figure 3 The Architecture of the AAM-Softmax Appearance Feature Descriptor Network

The Softmax classifier is widely used in deep classification tasks, with loss function as Eq. (1):

$$L_{\text{standard}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^{\mathrm{T}}x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^{\mathrm{T}}x_i + b_j}} \tag{1}$$

where $x_i \in \mathbb{R}^d$ is the input feature (due to the l_2 Normalisation layer in the descriptor, $\|x_i\|$ is equal to 1), y_i is the class label of x_i. W_j is the j-th column of the weight matrix $W \in \mathbb{R}^{d\times n}$ and $b_j \in \mathbb{R}^n$ is the bias. The target logit term is presented as $W_{y_i}^{\mathrm{T}}x_i + b_{y_i}$. N is the batch size and n is the number of classes. Based on the Softmax loss, the improvement can boost the ability to learning discriminative features effectively. Specifically, the bias b_j is set to 0, and l_2 normalisation is imposed on W_j ($\|W_j\| = 1$) to project it onto the unit sphere. So the term $W_j^{\mathrm{T}}x_i + b_j$ is equal to $W_j^{\mathrm{T}}x_i = \|W_j\|\|x_i\|\cos\theta_j = \cos\theta_j$, where θ_j is the angle between W_j and x_i. The l_2-normalised Softmax loss is presented as Eq. (2):

$$L_{l_2\text{-normalization}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{\sum_{j=1}^{n} e^{s\cos\theta_j}} \tag{2}$$

where $\cos\theta_{y_i}$ the is target logit, s is the feature scale hyper-parameter. Then by imposing an additive angular margin m to the target logit, the AAM-Softmax loss[44] formulation can be written as Eq. (3):

$$L_{\text{aam}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}} \tag{3}$$

The additive angular margin penalty makes the decision boundaries more stringent and separated, enhancing the similarity of intra-class features and the disparity of inter-class features simultaneously, which facilitates to improve the discriminative capability of features effectively. Since the AAM-Softmax loss only imposes an additive angular margin constraint in the angular space, it neither increases the structural complexity of the network nor the number of trainable parameters. When performing the online MOT task, the pre-trained descriptor network is exploited as the feature encoder to obtain discriminative features of pedestrians for the appearance distance metric in the subsequent data association process.

3.1.3 Assignment Problem

During the data association process, the assignment costs between tracklets and detections are the basis of association. In this paper, based on different states of the tracklets, different metric methods are utilised to calculate the assignment costs and hierarchical associations are conducted. For each tracklet, except the initial-state tracklet, an appearance feature gallery $\Phi_i = \{\boldsymbol{f}_k^{(i)}\}_{k=1}^{K}$will be generated, containing the historical appearance features backtracking from the current frame, where $\boldsymbol{f}_k^{(i)}(\|\boldsymbol{f}_k^{(i)}\|=1)$is the k-th l_2-normalised historical appearance feature, k=1 denotes the current frame and K is the maximum capacity of the gallery. The appearance metric between the tracklet and the detection forms an important part of the distance metric. It can be derived by calculating the cosine distance between all historical appearance features of the tracklets $\boldsymbol{f}_k^{(i)} \in \Phi_i$and those of the detections one by one and selecting the minimum value, written as Eq. (4):

$$d_{\text{a}}(t^{(i)}, d_j) = \min\{1 - \boldsymbol{f}_j^T \boldsymbol{f}_k^{(i)} \mid \boldsymbol{f}_k^{(i)} \in \Phi_i\} \tag{4}$$

where $t^{(i)}$is the i-th tracklet (except for the initial-state tracklet), d_j is the$j$-th detection and f_j is the appearance feature of d_j. $d_{\text{a}}(t^{(i)}, d_j)$ represents the appearance metric between $t^{(i)}$and d_j. Due to the l_2 nomalisation operation ($\|\boldsymbol{f}_k^{(i)}\|=1$, $\|\boldsymbol{f}_j\|=1$), the cosine similarity can be written as the inner product form $\boldsymbol{f}_j^{\text{T}}\boldsymbol{f}_k^{(i)}$.

① Assignment Cost Calculation Based on the State

The Initial-State: For the initial-state tracklet, since there is no appearance feature gallery, only position information can be used for data association. Firstly, the standard Kalman filter is used to predict its moving state. Then the inverse of the intersection-over-union (IOU) between the prediction bounding box and the detection bounding box, defined as the IOU distance, is calculated as the motion metrics, presented as Eq. (5):

$$d(t_{\text{initial}}^{(i)}, d_j) = \frac{1}{\text{IoU}(t_{\text{initial}}^{(i)}, d_j)} = \frac{\text{B}_{t_{\text{initial}}^{(i)}} \cup \text{B}_{d_j}}{\text{B}_{t_{\text{initial}}^{(i)}} \cap \text{B}_{d_j}} \tag{5}$$

where $t_{\text{initial}}^{(i)}$ denotes the i-th initial-state tracklet and $\text{B}_{t_{\text{initial}}^{(i)}}$ is the prediction bounding box of $t_{\text{initial}}^{(i)}$; d_j denotes the j-th detection and B_{d_j} is the bounding box of d_j . $d(t_{\text{initial}}^{(i)}, d_j)$ and $\text{IoU}(t_{\text{initial}}^{(i)}, d_j)$ represent the distance metrics and IoU value between $t_{\text{initial}}^{(i)}$ and d_j .

The Tentative-state: The tentative-state tracklet already generates an appearance feature gallery Φ_i , but at this time the features in the gallery are limited. So, at this stage, the IOU distance between the prediction and the detection is still used as the motion metric for association, whilst the appearance metric is only used as a threshold for filtering and discarding infeasible detections. If the appearance metric is greater than the threshold, the candidate detection will be excluded, with no possibility of being associated. The assignment cost is expressed as Eq. (6):

$$d(t_{\text{tentative}}^{(i)}, d_j) = \begin{cases} \dfrac{1}{\text{IoU}(t_{\text{tentative}}^{(i)}, d_j)} & d_{\text{a}}(t_{\text{tentative}}^{(i)}, d_j) <= \tau_{\text{a}} \\ \infty & \text{else} \end{cases} \tag{6}$$

where $t_{\text{tentative}}^{(i)}$ and d_j denote the i-th tentative-state tracklet and the j-th detection respectively. $d(t_{\text{tentative}}^{(i)}, d_j)$, $d_{\text{a}}(t_{\text{tentative}}^{(i)}, d_j)$ and $\text{IoU}(t_{\text{tentative}}^{(i)}, d_j)$ represent the metric distance, the appearance metric (obtained by Eq. (4)) and the IoU value between $t_{\text{tentative}}^{(i)}$ and d_j respectively. τ_{a} is the appearance threshold.

The Stable-State: For the stable-state tracklet, the data association process contains two stages. In the first association stage, the Mahalanobis distance is utilised as the motion metrics, written as Eq. (7)

$$d_{\text{m}}(t_{\text{stable}}^{(i)}, d_j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i) \tag{7}$$

where $t_{\text{stable}}^{(i)}$ and d_j represent the i-th stable-state tracklet and the j-th detection respectively. $d_{\text{m}}(t_{\text{stable}}^{(i)}, d_j)$ is the Mahalanobis distance between $t_{\text{stable}}^{(i)}$ and d_j , and (y_i, S_i) is the measurement space of $t_{\text{stable}}^{(i)}$. In this stage, different metric methods are adopted for different states of the camera (moving or stationary) due to the motion information caused by the camera's movement. For the moving situation, the appearance metrics is mainly considered, and the motion metrics is only used for filtering. If the motion metrics is greater than the threshold, the candidate detection will be discarded.The assignment cost is written as Eq. (8):

$$d(t_{\text{stable}}^{(i)}, d_j) = \begin{cases} d_{\text{a}}(t_{\text{stable}}^{(i)}, d_j) & d_{\text{m}}(t_{\text{stable}}^{(i)}, d_j) <= \tau_{\text{m}} \\ \infty & \text{else} \end{cases} \tag{8}$$

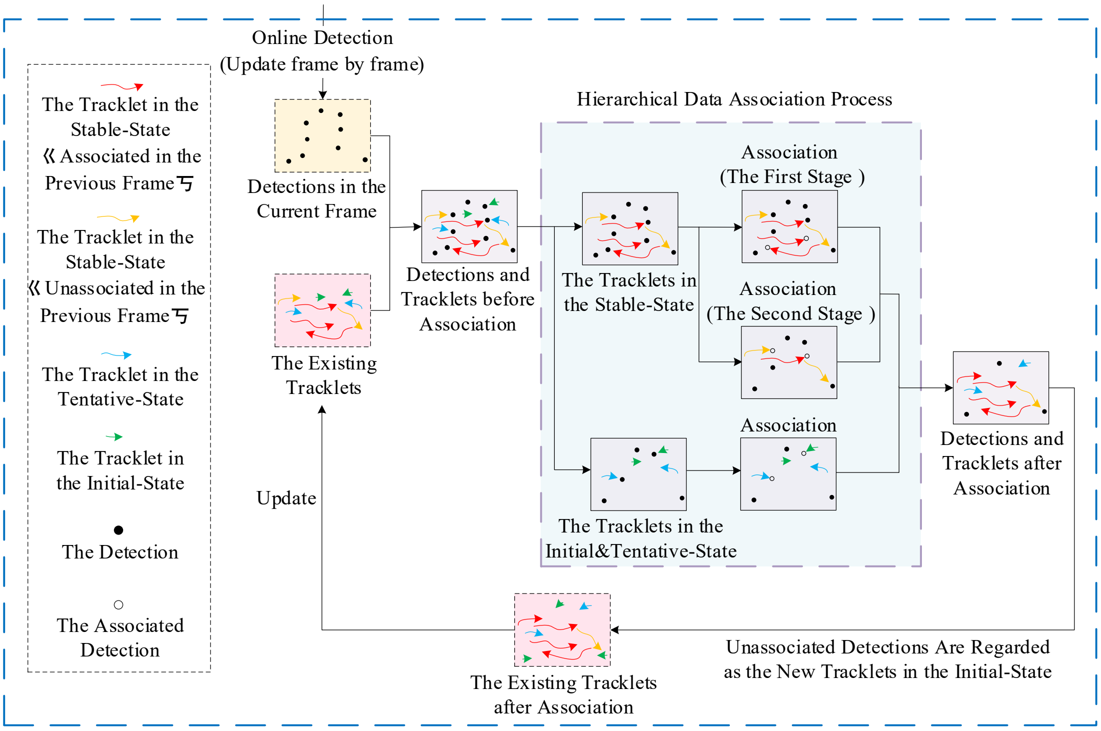
where $d(t_{\text{stable}}^{(i)}, d_j)$, $d_{\text{a}}(t_{\text{stable}}^{(i)}, d_j)$ (calculated by Eq.(4)) and $d_{\text{m}}(t_{\text{stable}}^{(i)}, d_j)$ (calculated by Eq.(7)) denote the metirc distance, the appearance metrics and the motion metrics between $t_{\text{stable}}^{(i)}$ and d_j respectively. τ_{m} is the Mahalanobis threshold. For the stationary situation, the motion metrics and the appearance metrics are both taken as the joint metrics, which are integrated into a unified form through the hyper-parameter λ, presented as Eq. (9):

$$d(t_{\text{stable}}^{(i)}, d_j) = \lambda d_{\text{m}}(t_{\text{stable}}^{(i)}, d_j) + (1-\lambda) d_{\text{a}}(t_{\text{stable}}^{(i)}, d_j) \tag{9}$$

Due to occlusions or other reasons, the appearance features of the object may change dramatically, causing the stable-state tracklet to fail to be associated in the first association stage. Hence, the second association stage is added for this situation. Instead of considering the appearance metrics, only the motion metrics calculated by IoU distance is utilised for data association in the second stage.

② Hierarchical Data Association and the States' Update

The stability of the tracklet determines its confidence. Therefore, according to the order of confidence from high to low, the corresponding states are the stable-state, the tentative state and the initial state, respectively. Based on this confidence order, a hierarchical association method is designed to divide the entire data association stage into three levels. The flow chart of the proposed hierarchical data association is shown in Figure 4. Specifically, the first association of stable-state tracklets is performed. Then, those tracklets that fail to be associated subsequently enter the second association stage. After that, the tracklets in the initial-state or the tentative-state are considered to be associated with the remaining detections. According to the state of the tracklet, the corresponding metrics is calculated, and the association between the tracklets and the detections is conducted by using the Hungarian algorithm [42]. For the tracklet in the initial-state or the tentative-state, if it fails to be associated, it will be deleted. This way of dealing with unstable objects can facilitate the tracker to filter the incorrectly detected objects to a certain extent thus to improve the performance efficiency. For the tracklet in the stable-state, if it fails to be associated for over consecutive τ_2 frames, it will be deleted, which can increase the completeness of the trajectory. For unassociated detection, it will be considered as a new tracklet and will be assigned with a new ID.



Online Detection
(Update frame by frame)
The Tracklet in the Stable-State 《Associated in the Previous Frame丂
The Tracklet in the Stable-State 《Unassociated in the Previous Frame丂
The Tracklet in the Tentative-State
The Tracklet in the Initial-State
The Detection
The Associated Detection
Detections in the Current Frame
The Existing Tracklets
Detections and Tracklets before Association
Hierarchical Data Association Process
Association (The First Stage)
The Tracklets in the Stable-State
Association (The Second Stage)
Association
The Tracklets in the Initial&Tentative-State
Detections and Tracklets after Association
Update
The Existing Tracklets after Association
Unassociated Detections Are Regarded as the New Tracklets in the Initial-State

Figure 4 The Flow Chart of the Proposed Hierarchical Data Association

The description of a tracklet includes the ID, the position information (the coordinates of the bounding box), the appearance feature and its state. After each association, the tracklets will be updated. For the successfully associated tracklet, the coordinates of the tracking bounding box are updated according to its new position; the new appearance feature in the current frame will be added to the appearance feature gallery. But when occlusions occur, the confidence of the appearance feature decreases due to the introduced noise. To tackle this issue, the IoU values between the prediction bounding box of the tracklet with all detection bounding boxes are calculated and a threshold τ_{IoU} is set. If there exists an IoU value greater than τ_{IoU}, the appearance feature gallery is not updated. Besides, the state of the tracklet should be updated as well according to its current state and the transition mechanism (Section 3.1.1).

As a summary, the entire process of online multi-object tracking based on hierarchical data association is illustrated in Figure 5. For the k-th frame, firstly the detector is used to conduct multi-object detection. Then, according to different states of the tracklets, different assignment cost calculation methods are adopted to associate the tracklets with the detections hierarchically. After data association, it is required to update the appearance feature galleries and the states of the tracklets. Then the updated tracklets will be associated with the candidate detections in the k+1-th frame.

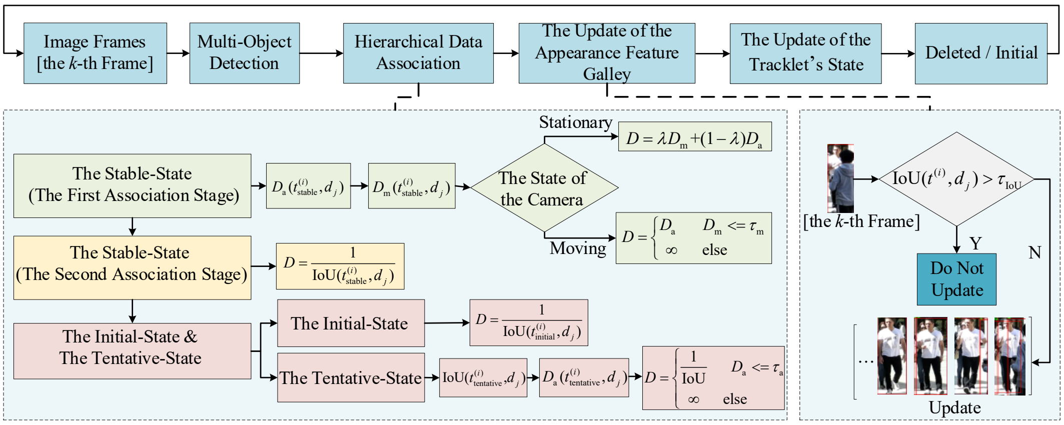


Image Frames [the k-th Frame]
Multi-Object Detection
Hierarchical Data Association
The Update of the Appearance Feature Galley
The Update of the Tracklet's State
Deleted / Initial
The Stable-State (The First Association Stage)
$D_a(t^{(i)}_{stable}, d_j)$
$D_m(t^{(i)}_{stable}, d_j)$
The State of the Camera
Stationary
$D = \lambda D_m + (1-\lambda) D_a$
Moving
$D = \begin{cases} D_a & D_m <= \tau_m \\ \infty & \text{else} \end{cases}$
The Stable-State (The Second Association Stage)
$D = \frac{1}{\mathrm{IoU}(t^{(i)}_{stable}, d_j)}$
The Initial-State & The Tentative-State
The Initial-State
$D = \frac{1}{\mathrm{IoU}(t^{(i)}_{initial}, d_j)}$
The Tentative-State
$\mathrm{IoU}(t^{(i)}_{tentative}, d_j)$
$D_a(t^{(i)}_{tentative}, d_j)$
$D = \begin{cases} \frac{1}{\mathrm{IoU}} & D_a <= \tau_a \\ \infty & \text{else} \end{cases}$
[the k-th Frame]
$\mathrm{IoU}(t^{(i)}, d_j) > \tau_{IoU}$
Y
N
Do Not Update
...
Update

Figure 5 The Entire Process of Online Multi-Object Tracking Based on Hierarchical Data Association

3.2 Trajectory-Based Social Distancing Measurement and Analysis

The surveillance videos are taken from arbitrary perspective views, it is necessary to transform the original perspective videos into the bird's eye view to perform distance measurement. This is carried out by utilising the perspective transformation matrix. As the calibration is conducted for the transformation of the ground plane, the bottom-centre point of tracking bounding box of every trajectory in each frame is transformed into the bird's eye view as the sampling point of the trajectory. Then, the re-parameterization time information is added to ensure *t* cannot be backtracked, and the spatio-temporal trajectories of pedestrians are represented in the three-dimensional coordinates space (x', $y'$and *t*). For addressing the VSD problem, the discrete Fréchet distance [48] is utilised to measure the distance between each spatio-temporal trajectory pair. Finally, the social distancing between pedestrians in the real world can be estimated by multiplying the metric distance with the scaling factor.

3.2.1 Trajectory Transformation and Distance Metrics

The essence of calibration is to map the original video into the bird's eye view by performing the perspective transformation.

The formula of the perspective transformation is presented as Eq. (10):

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{M} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{10}$$

where (u, v) and $(u, v, 1)$ are the Cartesian coordinate and the homogeneous coordinate of every trajectory respectively in each original frame. $\mathbf{M}$ is the perspective transformation matrix.(x, y, z) is the calibrated homogeneous coordinate in the bird's eye frame. Its corresponding Cartesian coordinate (x', y') can be obtained as Eq. (11):

$$\begin{cases} x' = \frac{x}{z} = \frac{a_{11}u + a_{12}v + a_{13}}{a_{31}u + a_{32}v + a_{33}} \\ y' = \frac{y}{z} = \frac{a_{21}u + a_{22}v + a_{23}}{a_{31}u + a_{32}v + a_{33}} \end{cases} \tag{11}$$

For calibration, we first need to select a rectangular reference area in the shooting scene. Due to the arbitrary angle of the camera, the reference area appears as a quadrilateral in the original perspective view. Since the video is captured by a single camera (monocular camera), the calibration method is to map the quadrilateral in the original video to the bird's eye view to re-form a rectangle. Using the four pairs of vertex coordinates of the quadrilateral and the rectangle, the perspective transformation matrix $\mathbf{M}$ can be calculated by Eq. (10). Then, the mapping coordinates in the bird's eye view of each trajectory's sampling points can be calculated through Eq. (11).

The Fréchet distance [48] is to determine the distance between each spatio-temporal trajectory pairs P and Q in the calibrated space S by taking into account location and time ordering. Fréchet distance is defined as Eq. (12):

$$D(P,Q) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \{d(P(\alpha(t)), Q(\beta(t)))\} \tag{12}$$

where d is the distance function of the space S. $P(\alpha(t))$ and $Q(\beta(t))$ represent the spatial position of the P and Q at time t respectively. α and β are continuous and non-decreasing reparameterization. The discrete Fréchet distance [48] is an approximation of the continuous Fréchet distance. Firstly, the two trajectory curves P and Q are discretized and are represented as the sequences with p and q sampling points, namely the specific positions of the trajectories in each frame, presented with $\sigma(P) = P(p_1, \cdots, p_p)$ and $\sigma(Q) = Q(q_1, \cdots, q_q)$ respectively. Because the distance metrics is performed in the bird's eye view, the corresponding transformed sequences $\sigma(P') = P'(p'_1, \cdots, p'_p)$ and $\sigma(Q') = Q'(q'_1, \cdots, q'_q)$ need to be obtained by Eq. (11). A coupling L' between P' and Q' is a sequence of distinct pairs from $\sigma(P')$ and $\sigma(Q')$, written as Eq. (13):

$$L' = (p'_{a_1}, q'_{b_1}), (p'_{a_2}, q'_{b_2}), \cdots, (p'_{a_m}, q'_{b_m}) \tag{13}$$

where $a_1 = 1$ and $b_1 = 1$, $a_m = p$ and $b_m = q$, and for all $i = 1, \cdots, q$, we have $a_{i+1} = a_i$ or $a_{i+1} = a_i + 1$, $b_{i+1} = b_i$ or $b_{i+1} = b_i + 1$. The length $\|L'\|$ of the coupling L' is the length of the longest link in L':

$$\|L'\| = \max_{i=1,\cdots m} d(p'_{a_i}, q'_{b_i}) \tag{14}$$

We use Euclidean distance to calculate $d(p'_{a_i}, q'_{b_i})$. The discrete Fréchet distance between P' and Q' in the bird's eye view is defined as Eq. (15):

$$\delta_{dF}(P', Q') = \min \|L'\| \tag{15}$$

By Multipling the metric distance in the bird's eye view by the scaling factor *s*, the social distancing in the real world can be estimated. On the one hand, the Euclidean distance is used to measure the distance between each sampling point pairs of trajectories from a local perspective, presented as Eq. (16):

$$D_s = s \cdot d(p'_{a_i}, q'_{b_i}) \tag{16}$$

On the other hand, from a holistic view, the discrete Fréchet distance is exploited to measure the distance between trajectory pairs, presented as Eq. (17):

$$D_t = s \cdot \delta_{dF}(P', Q') \tag{17}$$

3.2.2 Social Distancing Analysis

Here we design a multi-scale social distancing analysis scheme to evaluate the social distancing situations in public spaces from multiple time scales. The scheme includes the following four evaluation metrics:

The Average Ratio of Pedestrians with Unsafe Social Distancing (ARP-USD): If the distance between pedestrians, calculated by Eq. (16), is below the minimum acceptable distance, we believe that the pedestrians are at an unsafe distance at this moment. The ARP-USD metric is the mean proportion of the number of pedestrians with an unsafe distance in respect to the total number of people over a period of time in the public space. Given a video with *M* frames, for the i-th frame, N_i is the total number of tracking persons, $p_i^k (k = 1, 2, \cdots, N_i)$ is the position point of the k-th pedestrian. The set of pedestrians with unsafe social distancing in the i-th frame is represented as Eq. (18):

$$T_i = \{p_i^k | s \cdot d(p_i'^k, p_i'^l) < \tau_s\} \tag{18}$$

where $p_i'^k, p_i'^l$ are the mapping points of $p_i^k, p_i^l (k, l = 1, 2, \cdots, N_i$ and $k \neq l)$ in the bird's eye view, $d(\bullet)$ is the Euclidean distance, s is the scaling factor, τ_s is the safe distance threshold, and n_i is the number of the elements in T_i. If $N_i \neq 0$, the ratio of pedestrians with an unsafe distance in the i-th frame can be written as n_i / N_i, so the ARP-USD is calculated as Eq. (19):

$$R_{\text{ARP-USD}} = \frac{1}{m}\sum_{i=1}^{m}\frac{n_i}{N_i} \tag{19}$$

where $m\ (m \leq M)$ is the number of the frames with $N_i \neq 0$.

The Number of Trajectory Pairs with Unsafe Social Distancing (NTP-USD): If the distance between the trajectory pair, calculated by Eq. (17), is below the safe distance threshold τ_s, we consider the trajectory pair to be at an unsafe distance. The NTP-USD metric is the number of the stable-state trajectory pairs with unsafe social distancing. Assuming that the number of the stable-state trajectories is N_S (N_S is dynamically updated) and the p-th stable-state trajectory is represented as $T_{\text{stable}}^{p}(p = 1, 2\cdots, N_S)$, the set of the stable-state trajectory pairs with unsafe social distancing can be formulated as Eq. (20):

$$Q=\{(T_{\text{stable}}^{p}, T_{\text{stable}}^{q}) | s \cdot \delta_{dF}(T_{\text{stable}}^{\prime p}, T_{\text{stable}}^{\prime q}) < \tau_s\} \tag{20}$$

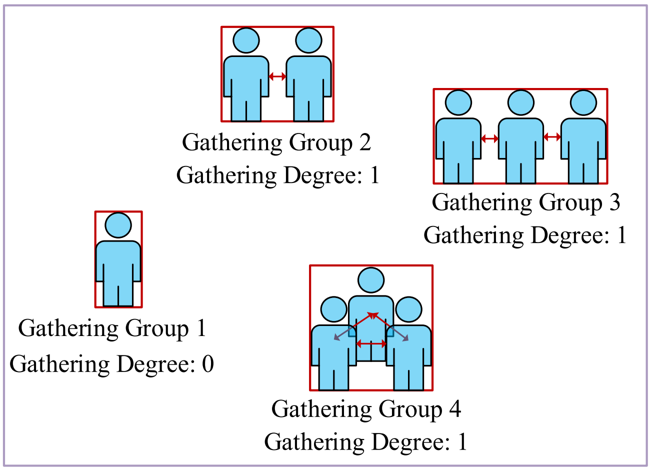
where $T_{\text{stable}}^{\prime p}, T_{\text{stable}}^{\prime q}$ are the mapped trajectories of $T_{\text{stable}}^{p}, T_{\text{stable}}^{q}(p, q = 1, 2\cdots, N_S \text{ and} p \neq q)$ in the bird's eye view, where $\delta_{dF}(\bullet)$ is the discrete Fréchet distance and s is the scaling factor. N_Q is the number of the elements in Q, indicating the value of the NTP-USD. If the Fréchet distance of the trajectory pair is less than the safe distance threshold, it means that the distance of each sampling point pair of the two trajectories has been less than the safe distance for the entire measurement process. In another word, the two pedestrians have continuely violated social distancing for a period of time.Therefore, based on spatio-temporal trajectories, the NTP-USD metrics measures the overall number of trajectory pairs with an unsafe distance in public spaces, which reflects the situation of social distancing violations in the public area over a period of time.

The Number of Pedestrian Pairs with Continuous Unsafe Social Distancing (NPPC-USD): Concerning the spread of pandemics, the longer people stay at an unsafe distance, the higher the risk of infection. Therefore, the duration of pedestrians staying within an unsafe distance is an essential factor that should be taken into account. The NTP-USD Fréchet distance is to measure the similarity of trajectories for a given duration holistically. But for two dissimilar trajectories, for example, two trajectories facing each other with opposite directions, their Fréchet distance can be very large. Hence NTP-USD is unable to spot the infection risk when pedestrians are very close for a certain amount of time but facing back to each other. So, for a trajectory pair, if the number of their sampling point pairs with an unsafe distance is more than the threshold τ_n, it is considered as the pedestrian pair with continuous unsafe social distancing. The NPPC-USD is designed to count the number of the above pedestrian pairs, which can reflect the concept of the unsafe distance for a period of time.

The Average Gathering Degree (AGD): To describe the degree of pedestrians gathering in public spaces, the concept of Gathering Group is defined. As shown in Figure 6, pedestrians in one gathering group have social distancing with one or more people in the same group less than the safe distance. The social distancing between any two people in different Gathering Groups is larger than the safe distance. According to the number of pedestrians in a group, the Gathering degree is divided into six levels, from 0 to 5 (illustratied in Figure 6). In order to facilitate unified grading, a single person is also regarded as a Gathering Group with Gathering degree 0. For each frame, the maximum Gathering degree is taken as the Gathering degree of this frame D_i. The Average Gathering degree of a video with M frames is formulated as Eq. (21):

$$D_{\mathrm{AGD}} = \frac{1}{M}\sum_{i=1}^{M} D_i \tag{21}$$

During the pandemic, the larger the number of people in the Gathering Group, the higher the risk of cross-infection. Therefore, the AGD can reflect the gathering situation of people in a period of time, which is a useful metric for assessing the risk of infection in public spaces.



The Number of the Gathering Group	Gathering Degree
1	0
2-6	1
7-12	2
13-20	3
21-30	4
More than 30	5

The Gathering Degree of This Frame: 1

Figure 6 The Gathering Group and Corresponding Gathering degree

4. Experiments and Discussion

4.1 Datasets

MOT16: We evaluate the performance of the proposed tracking method on the MOT16 benchmark dataset [49]. It consists of 14 video sequences, where 7 videos are used as training and verification sets, and another 7 are employed as test sets. The input sizes of the MOT16 are1920×1080 and 640×480. There are front-view scenes taken from moving camera and top-down view scenes captured from surveillance camera. The complex scenes, the large number of pedestrians and the varying laminations have imposed great challenges in analysing this MOT16 benchmark dataset.

SCU-VSD: we conduct social distancing measurment and analysis experiments on our own datasets, called as SCU-VSD. It includes 8 pedestrian video sequences, which were taken from a pedestrian street with different scenes and perspective views. For each video sequence, the size is 1920×1080, the duration is 60 seconds, and the frame rate is 25 fps (each video gives 1500 consecutive frames).

4.2 Implementation Details

4.2.1 Hierarchical Association Based Online Multi-Pedestrian Tracking

The detection results of the MOT16 benchmark used in the paper are provided by the POI method [40]. The detector in the POI is Faster R-CNN [25] fine-tuned by additional training datasets (including ETHZ pedestrian dataset [50], Caltech pedestrian dataset [51] and their surveillance dataset [40]). The AAM-Softmax appearance feature descriptor is trained using a large-scale person re-ID dataset Market-1501 [46] captured by six cameras. It contains 12,936 images of 751 identities for training, 3,368 images of another 750 identities as the query set, and 19,732 images as the gallery set. The input images are resized to 128×64. For the AAM-Softmax loss, the hyper-parameters s and m in Eq. (3) are 30 and 0.006 respectively. The Optimizer is Adam and the batch size is 128. During the training, the learning rate is set to 1×10^{-3} with the first 55,000 interactions, and it decays to 1×10^{-4} in the last 10,000 interactions. For the transition mechanism of the states, the thresholds τ_1 and τ_2 are set to 3 and 30 respectively. For the data association, the appearance threshold τ_{a} in Eq. (6) is set to 0.8 and Mahalanobis threshold τ_{m} in Eq. (8) is set to 9.49, the hyper-parameter λ in Eq. (9) is set to 0.2. For the states' update, the threshold τ_{IoU} is set to 0.5.

4.2.2 Trajectory-Based Social Distancing Measurement and Analysis

For each scene, a rectangular reference area on the ground is selected and its actual length and width are measured. Due to the arbitrary angle of the camera, the rectangular reference area is presented as a quadrilateral in the original perspective video. According to the aspect ratio of the reference area, a reference rectangle is drawn with scaling factor s = 0.1 in the bird's eye view (500×500), which corresponds to the calibrated rectangle of the quadrilateral in the original video. Through the coordinates of the four vertex pairs of the quadrilateral and the calibrated rectangle, the perspective transformation matrix **M** of each video can be obtained by using the Eq. (10). By using **M**, the transformed trajectory of each pedestrian in the bird's eye view can be calculated. The information of the selected rectangular reference area of each SCU-VSD video is shown in Table 1.

Table 1 The Information of the Selected Rectangular Reference Areas for SCU-VSD Videos

SCU-VSD		01	02	03	04	05	06	07	08
In Real World	Width (m)	6	6	7.8	6	6	6	6	7.2
	Length (m)	14.7	12	11	24.5	18	12	15	12.6
In Bird's Eye View	Width (pixel)	60	60	78	60	60	60	60	72
	Length (pixel)	147	120	110	245	180	120	150	126

For social distancing measurement, the safe distance threshold τ_s is set to 2 m, the threshold τ_n is set to 250, and the scaling factor s is set to 0.1. Based on the varying scales of time, the experiments of multi-scale social distancing measurement and analysis in public spaces are performed as follow: (1) for 1/25 second (one frame as a unit), the Euclidean distance between tracking objects in the bird's eye view is measured to calculate the real-time social distancing between pedestrians, the real-time ratio of pedestrians with unsafe social distancing and the real-time gathering degree; (2) for 10 seconds (250 consecutive frames as a unit), the ARP-USD and AGD are calculated; (3) for 60 seconds (1500 frames, this is the entire video duration), the ARP-

USD and AGD are calculated; for pedestrians' trajectories, the NTP-USD and the NPP-CUSD are calculated.

4.3 Experiments Results and Discussion

4.3.1 Hierarchical Data Association Based Online Multi-Pedestrian Tracking

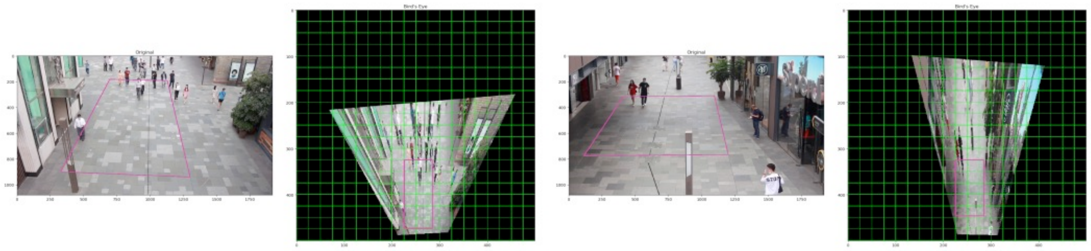
The results shown in Table 2 are obtained based on MOT16 Benchmark dataset. Compared with other online MOT methods, our proposed hierarchical data association based multi-pedestrian tracking method has achieved overall advanced performance. While maintaining high tracking accuracy and precision, the IDS of our proposed method decreases to 710, which effectively reduces the number of trajectory ID switches and improves the ability to maintain trajectory ID. The reduction of FM indicates the decrement of the number of trajectory interruptions.

Table 2 Comparisons of Different Online Algorithms on MOT16 Benchmark (with Private Detectors)

Method	MOTA (%) ↑	MOTP (%) ↑	IDS ↓	FM ↓	MT (%) ↑	ML (%) ↓
Config-MOT[30]	43.9	76.0	1030	**1795**	17.4	30.2
MOTDT[52]	47.6	50.9	792	—	15.2	**15.2**
STRN[53]	48.5	73.7	747	—	17.0	34.9
Deep Sort[32]	61.4	**79.1**	781	2008	32.8	18.2
EAMTT [28]	52.5	78.8	910	—	19.0	34.9
The proposed method	**61.4**	**79.1**	**710**	1913	**30.3**	19.9

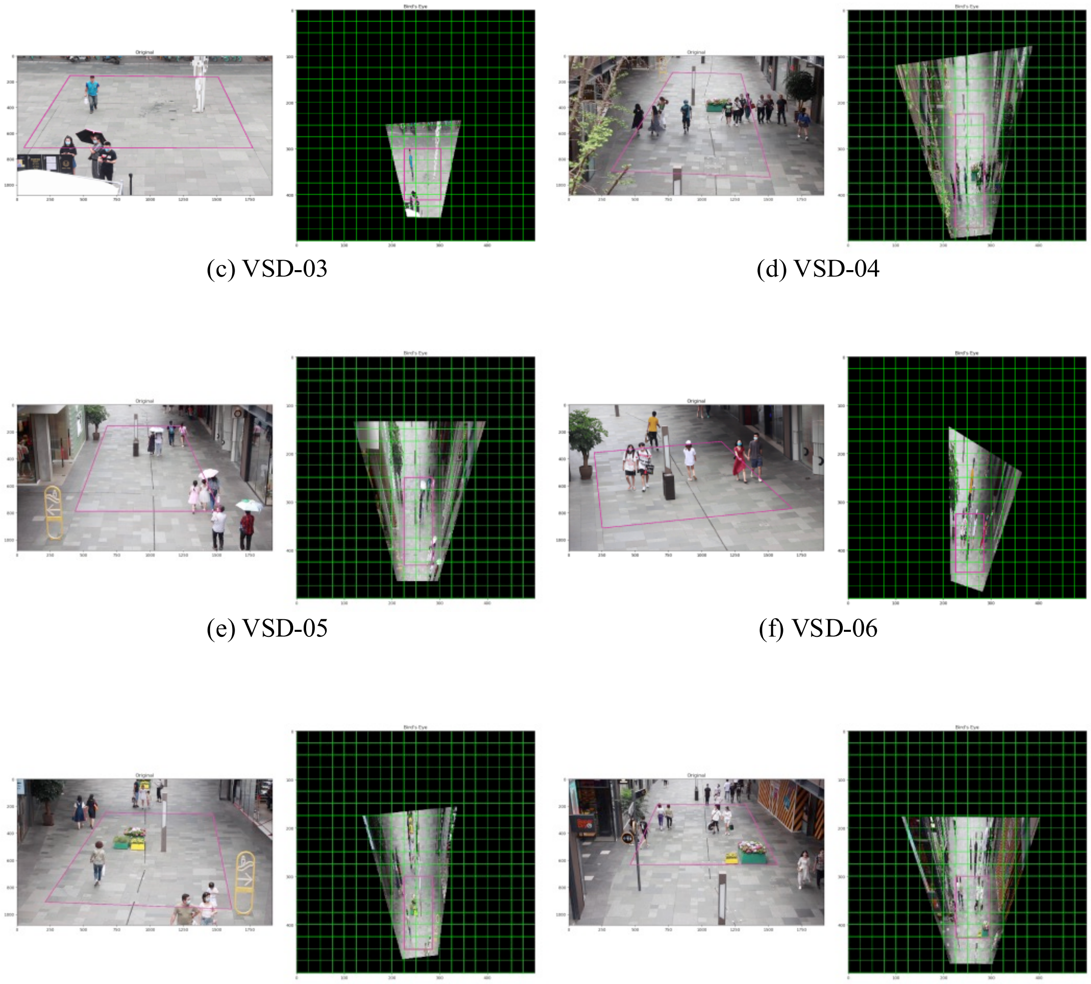
4.3.2 Trajectory-Based Social Distancing Measurement and Analysis

The comparisons between the original perspective view and the calibrated bird' s eye view for SCU-VSD dataset are shown in the figure 7. The rectangular reference area in each video is marked as a purple box. Due to the different perspective views of the videos, the reference areas in original videos are presented as different quadrilaterals.



(a) VSD-01

(b) VSD-02



(c) VSD-03

(d) VSD-04

(e) VSD-05

(f) VSD-06

(g) VSD-07

(h) VSD-08

Figure 7 The Comparisons between the Original Perspective View and the Calibrated Bird' s Eye View for SCU-VSD

In the calibration process, we use four corresponding vertex coordinate pairs of the reference area in the original video and the bird's eye view to calculate the perspective transformation matrix **M** of each video by Eq. (10), shown in Table 3. The numerical values of each matrix are expressed using scientific notation.

Table 3 The Coordinates of the Four Vertex Pairs and the Perspective Transformation Matrix for SCU-VSD

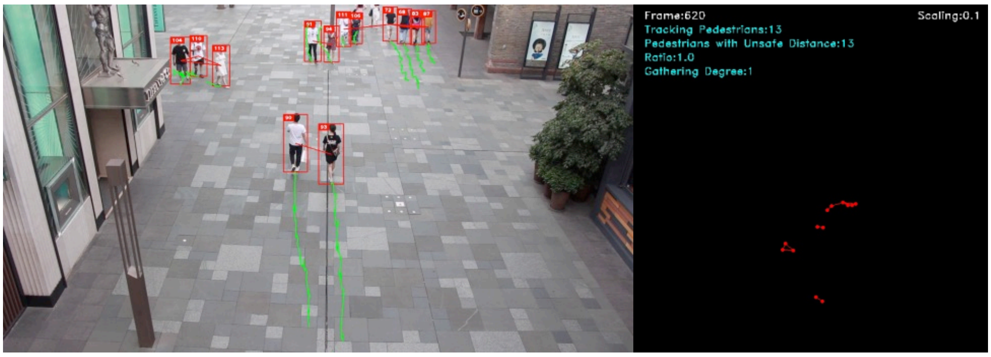
Dataset	in the original video	in bird's eye view	**M**
SCU-VSD-01	$P_1 = (695, 183)$, $P_2 = (328, 899)$ $P_3 = (1300, 940)$, $P_4 = (1129, 193)$	$P'_1 = (225, 325)$, $P'_2 = (225, 472)$ $P'_3 = (285, 472)$, $P'_4 = (285, 325)$	$\begin{bmatrix} 1.9422e-01 & 6.5468e-01 & 6.8420e+01 \\ -2.1163e-02 & 1.4486e+00 & 2.1647e+02 \\ -1.9038e-05 & 2.4575e-03 & 1.0000e+00 \end{bmatrix}$

SCU-VSD-02	$P_1 = (409, 317)$, $P_2 = (113, 776)$ $P_3 = (1199, 765)$, $P_4 = (1095,310)$	$P'_1 = (225, 325)$, $P'_2 = (225, 445)$ $P'_3 = (285, 445)$, $P'_4 = (285, 325)$	$\begin{bmatrix} 1.5449e-01 & 5.8471e-01 & 1.3333e+02 \\ 1.4504e-02 & 1.4124e+00 & 9.7908e+01 \\ 2.2426e-05 & 2.1704e-03 & 1.0000e+00 \end{bmatrix}$

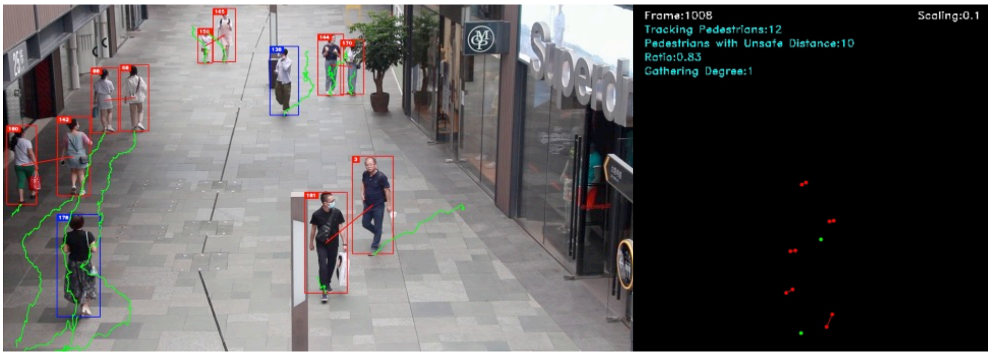
SCU-VSD-03	P1 = (397, 153), P2 = (51, 712)
	P3 = (1771, 1715), P4 = (1510, 163)
		P'1 = (225, 300), P'2 = (225, 410)
		P'3 = (303, 410), P'4 = (303, 300)
			$\begin{bmatrix} 7.4208e-02 & 3.0369e-01 & 1.8572e+02 \\ -1.0730e-02 & 6.9188e-01 & 2.4727e+02 \\ -2.5199e-05 & 1.1300e-03 & 1.0000e+00 \end{bmatrix}$

SCU-VSD-04	$P_1 = (781, 129)$, $P_2 = (320, 891)$, $P_3 = (1515, 940)$, $P_4 = (1295, 140)$	$P'_1 = (225, 225)$, $P'_2 = (225, 470)$, $P'_3 = (285, 470)$, $P'_4 = (285, 225)$	$\begin{bmatrix} 1.4485e-01 & 5.7949e-01 & 9.7719e+01 \\ -2.3251e-02 & 1.4215e+00 & 1.2039e+02 \\ -1.4727e-05 & 2.1771e-03 & 1.0000e+00 \end{bmatrix}$
SCU-VSD-05	$P_1 = (677, 155)$, $P_2 = (436, 788)$ $P_3 = (1529, 788)$, $P_4 = (1225, 155)$	$P'_1 = (225, 250)$, $P'_2 = (225, 430)$ $P'_3 = (285, 430)$, $P'_4 = (285, 250)$	$\begin{bmatrix} 1.4474e-01 & 5.2241e-01 & 1.1847e+02 \\ -8.9224e-16 & 1.2690e+00 & 1.3379e+02 \\ -2.1849e-18 & 2.0769e-03 & 1.0000e+00 \end{bmatrix}$
SCU-VSD-06	$P_1 = (190, 355)$, $P_2 = (250, 913)$ $P_3 = (1679, 768)$, $P_4 = (1146, 272)$	$P'_1 = (225, 325)$, $P'_2 = (225, 445)$, $P'_3 = (285, 445)$, $P'_4 = (285, 325)$	$\begin{bmatrix} 1.0968e-01 & 3.0204e-01 & 2.1112e+02 \\ 6.8670e-02 & 9.3750e-01 & 1.4407e+02 \\ 8.1190e-05 & 1.3861e-03 & 1.0000e+00 \end{bmatrix}$
SCU-VSD-07	$P_1 = (666, 247)$, $P_2 = (211, 909)$ $P_3 = (1613, 961)$, $P_4 = (1472, 263)$	$P'_1 = (225, 300)$, $P'_2 = (225, 450)$ $P'_3 = (285, 450)$, $P'_4 = (285, 300)$	$\begin{bmatrix} 1.0462e-01 & 4.1639e-01 & 1.3844e+02 \\ -9.2389e-03 & 9.9577e-01 & 1.7481e+02 \\ 4.6381e-06 & 1.5342e-03 & 1.0000e+00 \end{bmatrix}$
SCU-VSD-08	$P_1 = (665, 187)$, $P_2 = (461, 634)$ $P_3 = (1569, 631)$, $P_4 = (1328, 186)$	$P'_1 = (225, 300)$, $P'_2 = (225, 426)$ $P'_3 = (297, 426)$, $P'_4 = (297, 300)$	$\begin{bmatrix} 1.5073e-01 & 5.3939e-01 & 1.1180e+02 \\ 8.1498e-04 & 1.2834e+00 & 1.7667e+02 \\ -5.8150e-07 & 2.0913e-03 & 1.0000e+00 \end{bmatrix}$

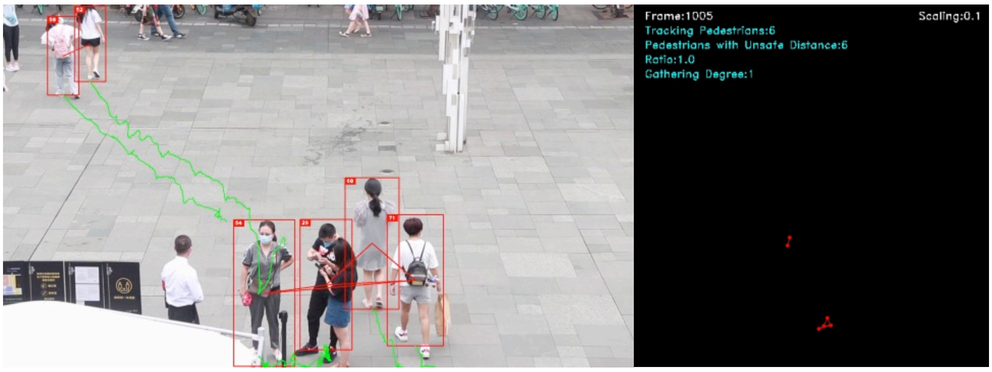
The real-time social distancing mesurement and analysis for SCU-VSD dataset are shown in Figure 8. The figure on the left is the original video, and the one on the right is the corresponding bird's eye view. The tracking pedestrians in the original video are transformed to trajectory points in the bird's eye view. The Euclidean distance between tracking object pairs in the bird's eye view are measured frame-by-frame to estimate the real-time social distances between the pedestrians. If the social distances between the pedestrian pair is less than the safe distance, the tracking bounding boxes in the left figure will change from blue to red, and the corresponding trajectory points in the right figure will change from green to red, with a red line linking pedestrians. The real-time ratio of pedestrians with unsafe social distancing and the real-time gathering degree are calculated, and the results are displayed in the top-left corner of the right figure.



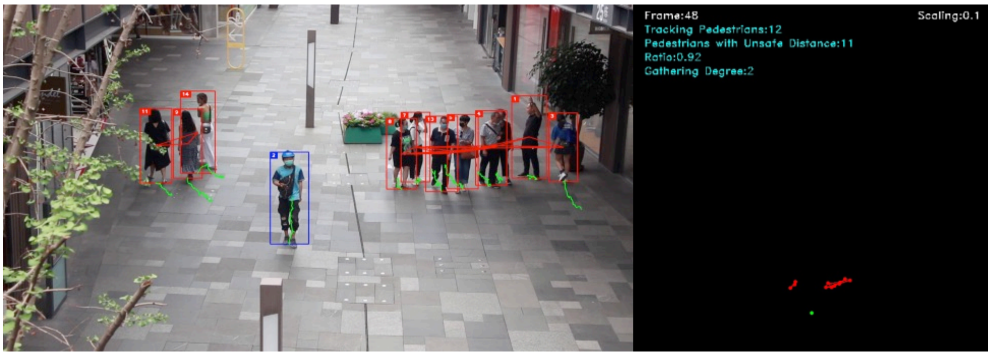
(a) SCU-VSD-01



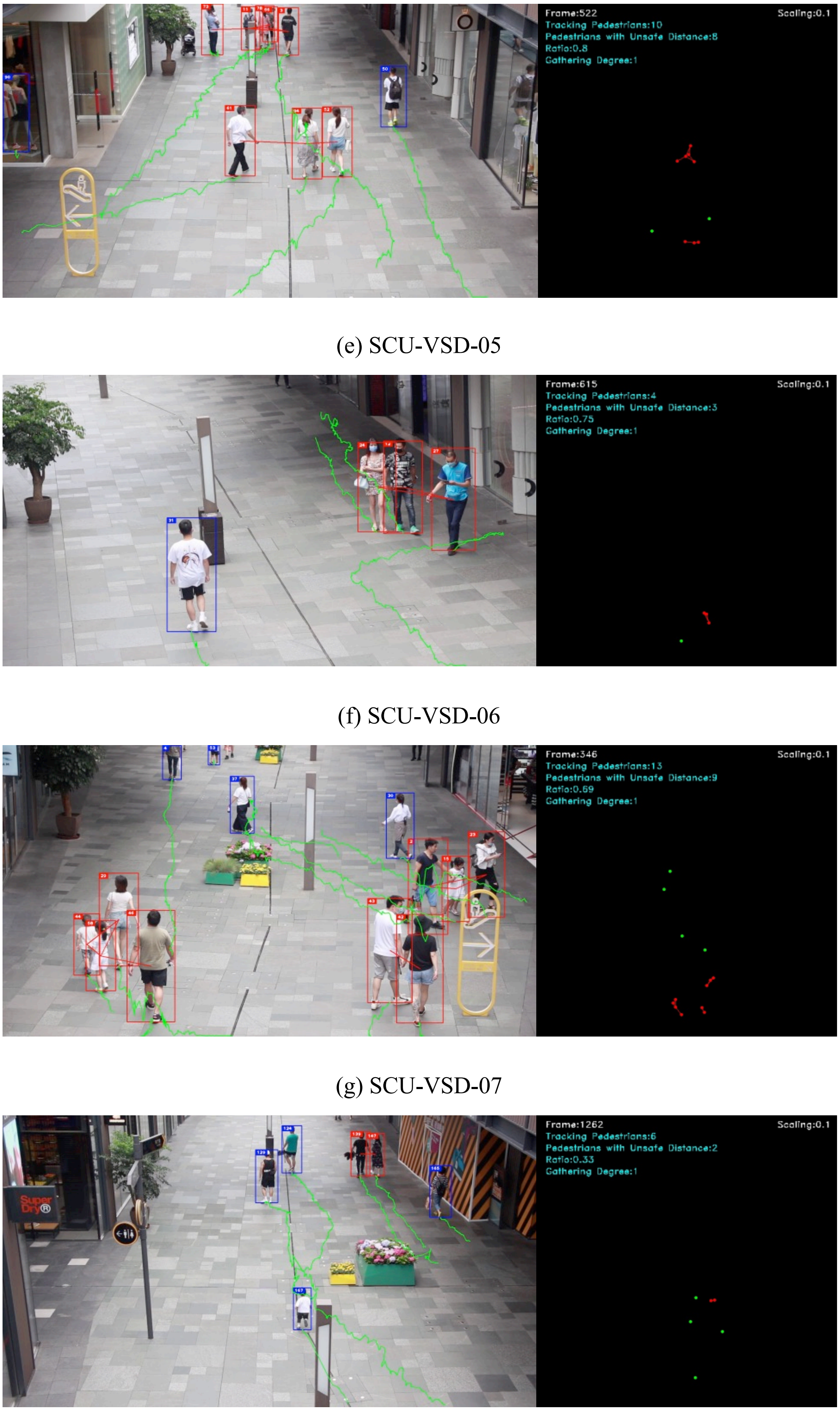
(b) SCU-VSD-02



(c) SCU-VSD-03



(d) SCU-VSD-04



(e) SCU-VSD-05

(f) SCU-VSD-06

(g) SCU-VSD-07

(h) SCU-VSD-08

Figure 8 The Real-Time Social Distancing measurement and Analysis for SCU-VSD Dataset

Taking 10s (250 consecutive frames) as a unit, each video is divided into 6 periods. For each period, the ARP-USD and AGD metrics are calculated, and the results are drawn using colormaps. The colormaps of ARP-USD (on the left) and AGD (on the right) for every 10s of each video clip are shown in Figure 9.

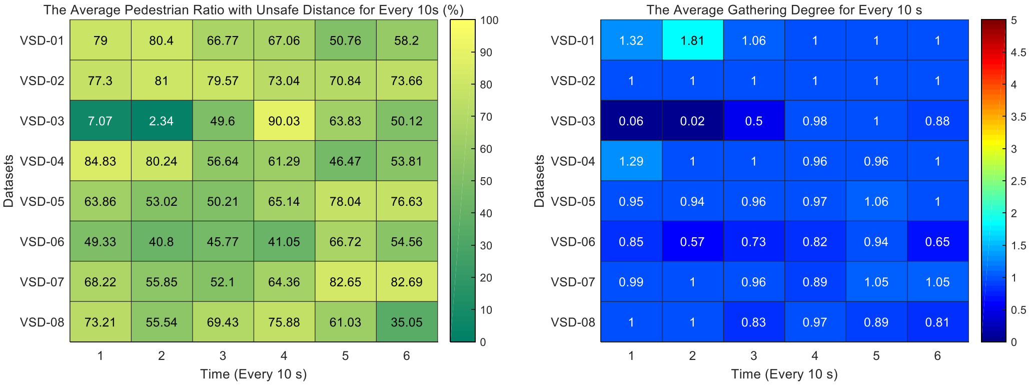


Figure 9 The Colormaps of ARP-USD and AGD for Every 10s

From Figure 9, each row of the colormaps can reflect the changing trend of metrics in different time units of the same video, while each column of the colormaps can reflect the metrics' changing tendency of different videos in the same time unit. These trends can be displayed intuitively through colour gradients of the corresponding colorbar. We take the first row and the first column of colourmaps as the examples for analysis. The first row of data represents the APR-USD and AGD metrics of VSD-01 clip in different time units. It can be seen that the changing trends of the two metrics of VSD-01 are roughly identical, rising and then falling, reaching their peaks (80.4% and 1.81 respectively) at the second time unit. The first column of the data represents the comparisons of the two metrics of the 8 video clips in the first time unit. It also can be observed that in this period, the two metric values of VSD-03 are minimum (7.07% and 0.06 respectively), while the APR-USD of VSD-04 and the AGD of VSD-01 are the maximum (84.83% and 1.32 respectively). In practical applications, the duration of the time window can be adjusted according to the actual requirement, so that APR-USD and AGD at different time scales can be obtained.

From the global perspective, the four metrics ARP-USD, NTP-USD, NPP-CUSD and AGD for each entire video are calculated, shown in Table 4.

Table 4 The Four Metrics ARP-USD, NTP-USD, NPP-CUSD and AGD for Each Entire video.

Metrics Datasets	ARP-USD (%)	NTP-USD	NPP-CUSD	AGD
SCU-VSD-01	67.01	32	11	1.2
SCU-VSD-02	75.90	10	9	1.0
SCU-VSD-03	47.62	3	2	0.57
SCU-VSD-04	63.84	22	11	1.03
SCU-VSD-05	64.48	14	10	0.98
SCU-VSD-06	49.59	5	4	0.76
SCU-VSD-07	67.64	19	6	0.99
SCU-VSD-08	61.60	6	6	0.92

As shown in Table 4, the NTP-USD, NPP-CUSD and AGD of SCU-VSD-01 achieve the maximum values, which are 32, 11 and 1.2 respectively, while the four metrics of SCU-VSD-03 are the minimum, 47.62%, 3, 2 and 0.57 respectively. Comprehensively, it can be concluded that SCU-VSD-01 video has the largest number of pedestrian pairs with unsafe social distancing and the highest average gathering degree. In contrast, SCU-VSD-03 video has the smallest number of pedestrians with unsafe distancing and the lowest average gathering degree.

5. Conclusion

In this paper, in response to the VSD problem in public places during the pandemic, we first proposed a hierarchical association based online multi-pedestrian tracking method to obtain pedestrians' trajectories. Then we proposed a spatio-temporal trajectory based multi-scale social distancing measurement and analysis method. The proposed VSD method considers both Euclidean distance from a static perspective and Fréchet distance from a spatio-temporal perspective to estimate the social distancing and analyse the crowd gathering situations based on a variety of time scales. The multi-scale metrics obtained by the proposed VSD approach can provide the local authorities with guiding information to help them monitor the real-time and overall situations of the social distancing of crowds in public spaces, so as to formulate and take corresponding prevention measures. In addition, for the areas where the pandemic has outbroken, the proposed VSD and analysis scheme can be used to provide useful supporting data for the subsequent epidemiological investigation, such as locating and search of the infection chain.

Reference

[1] World Health Organization (WHO) Emergency Committee. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). Geneva:WHO; 30 January 2020. Available from: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov)

[2] WHO Coronavirus Disease (COVID-19) Dashboard (Online). Available from: https://covid19.who.int/ (Accessed 6 January 2021).

[3] Aguilar J B, Faust J S, Westafer L M, et al. A model describing COVID-19 community transmission taking into account asymptomatic carriers and risk mitigation[J]. medRxiv, 2020.

[4] Dias S M, Queiroz K I P M, Martins A M. Controlling epidemic diseases based only on social distancing level[J]. arXiv preprint arXiv:2005.08052, 2020.

[5] Prem K, Liu Y, Russell T W, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study[J]. The Lancet Public Health, 2020.

[6] Hellewell J, Abbott S, Gimma A, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts[J]. The Lancet Global Health, 2020.

[7] Cacciapaglia G, Sannino F. Interplay of social distancing and border restrictions for pandemics (COVID-19) via the epidemic Renormalisation Group framework[J]. arXiv preprint arXiv:2005.04956, 2020.

[8] Sun C，Zhai Z. The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission [J]. Sustainable Cities and Society, 2020.102390.

[9] Rahmani A M, Mirmahaleh S Y H. Coronavirus disease (COVID-19) prevention and treatment methods and effective parameters: A systematic literature review[J]. Sustainable Cities and Society, 2020.102568.

[10] Silva B N , Khan M , Han K . Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities[J]. Sustainable Cities and Society, 2018, 38,

697–713.

[11] Bibri S E, Krogstie J. Smart sustainable cities of the future: An extensive interdisciplinary literature reviCew[J]. Sustainable Cities and Society, 2017, 31:183-212.

[12] He H, Li R, Wang R, et al. Efficient Suspected Infected Crowds Detection Based on Spatio-Temporal Trajectories[J]. arXiv preprint arXiv:2004.06653, 2020.

[13] Zhou C, Yuan W, Wang J, et al. Detecting Suspected Epidemic Cases Using Trajectory Big Data[J]. arXiv preprint arXiv:2004.00908, 2020.

[14] Silva J C S, de Lima Silva D F, Neto A S D, et al. A city cluster risk-based approach for Sars-CoV-2 and isolation barriers based on anonymized mobile phone users' location data[J]. Sustainable cities and society, 2020: 102574.

[15] Bhattacharya S, Maddikunta P K R, Pham Q V, et al. Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey[J]. Sustainable cities and society, 2020: 102589.

[16] Loey M, Manogaran G, Taha M H N, et al. Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection[J]. Sustainable Cities and Society, 2020: 102600.

[17] Nguyen C T, Saputra Y M, Van Huynh N, et al. Enabling and Emerging Technologies for Social Distancing: A Comprehensive Survey and Open Problems [J]. arXiv preprint arXiv:2005.02816, 2020.

[18] Cristani M, Del Bue A, Murino V, et al. The Visual Social Distancing Problem[J]. arXiv preprint arXiv:2005.04813, 2020.

[19] Khandelwal P, Khandelwal A, Agarwal S. Using Computer Vision to enhance Safety of Workforce in Manufacturing in a Post COVID World[J]. arXiv preprint arXiv:2005.05287, 2020.

[20] Yang D, Yurtsever E, Renganathan V, et al. A Vision-based Social Distance and Critical Density Detection System for COVID-19[J]. arXiv preprint arXiv:2007.03578, 2020.

[21] Shorfuzzaman M, Hossain M S, Alhamid M F. Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic[J]. Sustainable cities and society, 2020, 64: 102582.

[22] Punn N S, Sonbhadra S K, Agarwal S. Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques[J]. arXiv preprint arXiv:2005.01385, 2020.

[23] Sathyamoorthy A J, Patel U, Savle Y A, et al. COVID-robot: Monitoring social distancing constraints in crowded scenarios[J]. arXiv preprint arXiv:2008.06585, 2020.

[24] Ahmed I, Ahmad M, Rodrigues J J P C, et al. A deep learning-based social distance monitoring framework for COVID-19[J]. Sustainable Cities and Society, 2020: 102571.

[25] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[26] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[27] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[28] Sanchez-Matilla R, Poiesi F, Cavallaro A. Online multi-target tracking with strong and weak detections[C]//European Conference on Computer Vision. Springer, Cham, 2016: 84-99.

[29] Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]//2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016: 3464-3468.

[30] Bae S H, Yoon K J. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(3): 595-610.

[31] Chu Q, Ouyang W, Li H, et al. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4836-4845.

[32] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE international conference on image processing (ICIP). IEEE, 2017: 3645-3649.

[33] Al-Shakarji N M, Bunyak F, Seetharaman G, et al. Multi-object tracking cascade with multi-step data association and occlusion handling[C]//2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018: 1-6.

[34] Brendel W, Amer M, Todorovic S. Multiobject tracking as maximum weight independent set[C]//CVPR 2011. IEEE, 2011: 1273-1280.

[35] Milan A, Roth S, Schindler K. Continuous energy minimization for multitarget tracking[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(1): 58-72.

[36] Dehghan A, Modiri Assari S, Shah M. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4091-4099.

[37] Son J, Baek M, Cho M, et al. Multi-object tracking with quadruplet convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5620-5629.

[38] Reid D. An algorithm for tracking multiple targets[J]. IEEE Transactions on Automatic Control, 1979, 24(6): 843-854.

[39] Fortmann T, Barshalom Y, Scheffe M, et al. Sonar tracking of multiple targets using joint probabilistic data association[J]. IEEE Journal of Oceanic Engineering, 1983, 8(3): 173-184.

[40] Yu F, Li W, Li Q, et al. Poi: Multiple object tracking with high performance detection and appearance feature[C]//European Conference on Computer Vision. Springer, Cham, 2016: 36-42.

[41] Kalman R E. A new approach to linear filtering and prediction problems[J]. Journal of Basic Engineering, 1960 ,82 (Series D): 35–45.

[42] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval research logistics quarterly, 1955, 2(1-2): 83-97.

[43] Wojke N, Bewley A. Deep cosine metric learning for person re-identification[C]//2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018: 748-756.

[44] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4690-4699.

[45] Jie S U, He X, Qing L, et al. A New Discriminative Feature Learning for Person Re-Identification Using Additive Angular Margin Softmax Loss[C]//2019 UK/China Emerging Technologies (UCET). IEEE, 2019: 1-4.

[46] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1116-1124.

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

[48] Eiter T, Mannila H. Computing discrete Fréchet distance[R]. Technical Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria, 1994.

[49] Milan A, Lealtaixe L, Reid I D, et al. MOT16: A Benchmark for Multi-Object Tracking[J]. arXiv: Computer Vision and Pattern Recognition, 2016.

[50] Ess A, Leibe B, Schindler K, et al. A mobile vision system for robust multi-person tracking[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.

[51] Dollár P, Wojek C, Schiele B, et al. Pedestrian detection: A benchmark[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 304-311.

[52] Chen L, Ai H, Zhuang Z, et al. Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification[C]. international conference on multimedia and expo, 2018: 1-6.

[53] Xu J, Cao Y, Zhang Z, et al. Spatial-Temporal Relation Networks for Multi-Object Tracking[J]. arXiv: Computer Vision and Pattern Recognition, 2019.