Automatic Identification of Human Subgroups in Time-Dependent Pedestrian Flow Networks (Supplementary Materials)

Wenhan Wu, Wenfeng Yi, Jinghai Li, Maoyin Chen *Member, IEEE*, and Xiaoping Zheng

PEDESTRIAN TRAJECTORY EXTRACTION

Fig. S1 shows the extraction process of pedestrian trajectories in the self-build dataset. First, a surveillance camera installed on the exterior wall of the canteen is used to collect the pedestrian video with a frame rate of 25 FPS. Then, we combine YOLOv4 [1] with Deep SORT [2] to implement the detection and tracking of pedestrians, and eliminate the perspective distortion using homography transformation. Last, 12,326 pedestrian trajectories are obtained after the cleaning process, and the temporal series of density, speed, and flow are calculated for trajectory analysis. The detailed procedures of data extraction are presented in subsequent contents.

A. Video Collection

In September 2020, we conducted several field investigations at Bashu Secondary School (106.56°E and 29.54°N) in Chongqing, China. The flat ground in front of the canteen can be chosen as a study area S because dining is an essential activity for school personnel at specific periods of the day. The study area S is restricted to pedestrian walking, and its area is estimated as $A_S \approx 453.23\,\mathrm{m}^2$ by field measurements. The students from Bashu Secondary School are regarded as the main research subjects, which is attributed to the following two reasons: First, subgroups frequently appear in this area due to the familiarity of students. Second, the behavior of students entering and leaving the canteen is relatively simple, whereas other behaviors such as staying or wandering seldomly appear.

To collect the pedestrian video in this area, a surveillance camera HIKVISION DS-2CD3T45-I5 (6 mm) was installed on the exterior wall of the canteen, roughly 6 m above the ground. The frame rate is controlled to 25 FPS and the resolution is 1920×1080 pixels. Fig. S1(a) shows the realistic pictures of the study area and the camera location, as well as a satellite image of the observation conditions. The camera collected 24-hour video from 17:00 on September 9, 2020 to 17:00 on September 10, 2020. Note that the request concerning video collection was approved by Bashu Secondary School, and all pedestrians were observed under natural conditions without being informed in advance.

Wenhan Wu, Wenfeng Yi, Maoyin Chen, and Xiaoping Zheng are with the Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China. (e-mail: wwh19@mails.tsinghua.edu.cn; ywf19@mails.tsinghua.edu.cn; mychen@mail.tsinghua.edu.cn; asean@mail.tsinghua.edu.cn).

Jinghai Li is with the School of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, Beijing 100029, China (e-mail: ljhai725@163.com).

B. Detection and Tracking

With the development of computer vision, the detection-based tracking method has become a significant technique. Here, YOLOv4 and Deep SORT algorithms are combined to achieve the detection and tracking of pedestrians. YOLOv4, as a continuation version of YOLO [3] in one-stage object detectors, achieves faster and stronger object detection mainly by adding tricks in the training process, loss calculation, and activation functions. Deep SORT, as an improved version of SORT [4] in multiple objective tracking (MOT) algorithms, reduces a host of invalid ID switches by incorporating more reliable metrics of motion and appearance information.

YOLOv4 is adopted to detect pedestrians in the collected video, and the concrete details are provided below. First, the collected video is split frame by frame into images as the input. Second, the feature layers on frame images are extracted through the CSPDarknet53 backbone, and a series of stacked residual network structures improve the detection performance by adding appropriate depth. Next, SPP [5] and PAN [6] are served as the feature pyramid structures, where SPP separates out the most important contextual features without the loss of speed, and PAN realizes multi-channel feature fusion by shortening the propagation of feature information between low and high levels. Finally, the object prediction is performed on multiple extracted feature layers, and each detected pedestrian is marked by the bounding box (x, y, a, b) that includes the center position (x, y), aspect ratio a, and height b.

The bounding boxes of detected pedestrians are regarded as the input of Deep SORT, and the original four-dimensional state space (x, y, a, b) can be extended to eight-dimensional $(x, y, a, b, \dot{x}, \dot{y}, \dot{a}, \dot{b})$, where $(\dot{x}, \dot{y}, \dot{a}, \dot{b})$ is the inter-frame velocity of parameters. The pedestrian trajectory is predicted using the Kalman filtering containing a uniform velocity model and a linear observation model. The measurement association records the trajectory duration since the last successful match, and the trajectory will be terminated if the duration exceeds a predefined maximum threshold. To associate the Kalman states of predicted trajectory with actual measurement values, similarity metrics are applied to compare motion and appearance information, where the Mahalanobis distance is valid for calculating motion information, and the smallest cosine distance is used to measure appearance information. The above similarity metrics are combined with a weighted factor for constructing associations to solve assignment issues.

The predicted uncertainty of Kalman filtering increases

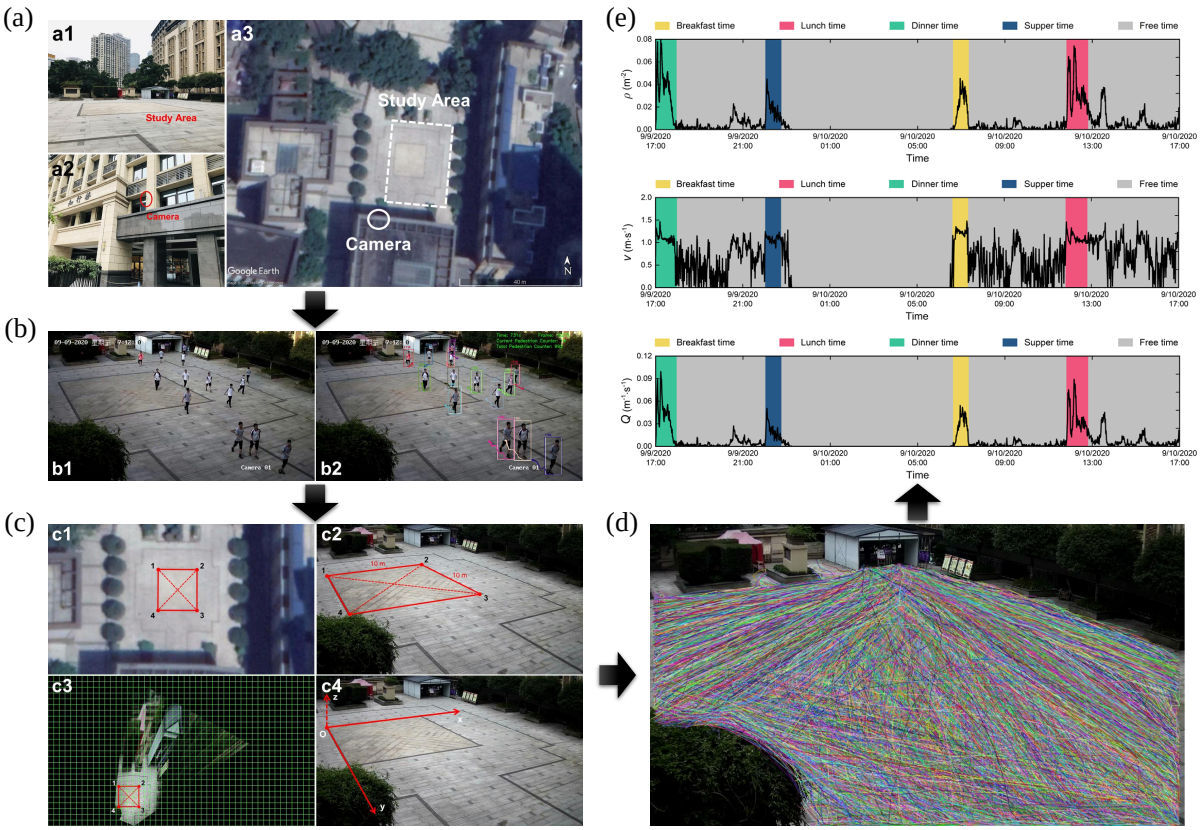

Fig. S1. Extraction process of pedestrian trajectories. (a) Realistic pictures of the study area (a1) and the camera location (a2), as well as a satellite image of the observation conditions (a3). (b) An original snapshot (b1) and its corresponding tracking snapshot (b2) at a certain time in the video sequence, where each pedestrian is marked with a bounding box and is assigned a unique ID number. (c) 4 pairs of feature matching points are selected from the satellite picture (c1), whose real-world coordinates and image coordinates are measured by fieldwork and software (c2). The frame images are calibrated for perspective distortion (c3), and a real-world coordinate system is established (c4). (d) 12,326 valid pedestrian trajectories are extracted after the cleaning process. (e) Temporal series of density, speed, and flow calculated based on the extracted trajectories. These five periods are highlighted by shaded areas with different colors, and the minimum sampling range is defined in minutes.

significantly if a pedestrian has been obscured for a long time. Assuming that two trajectories compete for the right to match the same detection, the one with longer occlusion is prone to be associated. This undesirable result would destroy the continuity of tracking because the Mahalanobis distance is conducive to larger uncertainty. Hence, the matching cascade algorithm gives priority to more frequent objects for solving this problem. In the following application phase, a feature extraction network of deep learning is used on large-scale person re-identification datasets, which is critical for distinguishing different pedestrians. Fig. S1(b) displays an original snapshot and its corresponding tracking snapshot at a certain time in the video sequence, where each pedestrian is marked with a bounding box and is assigned a unique ID number.

C. Calibration of Perspective Distortion

The frame images from the collected video are affected by perspective distortion, resulting in each pixel corresponding to a different metric size. Therefore, it is necessary to achieve the reconstruction of the coordinate system using the calibration of perspective distortion. To describe the geometric properties of perspective projection, the mapping function $(u_i, v_i) = f(X_i, Y_i, Z_i)$ is employed to convert the real-world coordinate (X_i, Y_i, Z_i) and the image coordinate (u_i, v_i). The calibration procedure of direct linear transformation (DLT) [7] is adopted to implement the mapping relation by ignoring nonlinear radial and tangential distortion components. Given that the positions of pedestrians are on the same plane (we assume $Z = 0$), the mapping relation is simplified below:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ 1 \end{bmatrix} \tag{S1}$$

where $\mathbf{H} = (h_{ij}) \in \mathcal{R}^{3\times3}$ denotes the invertible homography matrix, whose coefficients depend on the internal parameters of the camera and the relative position of the camera in space. To solve these coefficients in the homography matrix, the above equation is transformed into the following form:

$$u_i = \frac{h_{11}X_i + h_{12}Y_i + h_{13}}{h_{31}X_i + h_{32}Y_i + h_{33}}, v_i = \frac{h_{21}X_i + h_{22}Y_i + h_{23}}{h_{31}X_i + h_{32}Y_i + h_{33}} \tag{S2}$$

Assuming that we obtain N pairs of feature matching points between image coordinates and real-world coordinates, the

system of linear equations is given by:

$$\begin{bmatrix} X_1 & Y_1 & 1 & 0 & 0 & 0 & -u_1X_1 & -u_1Y_1 & -u_1 \\ 0 & 0 & 0 & X_1 & Y_1 & 1 & -v_1X_1 & -v_1Y_1 & -v_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N & Y_N & 1 & 0 & 0 & 0 & -u_NX_N & -u_NY_N & -u_N \\ 0 & 0 & 0 & X_N & Y_N & 1 & -v_NX_N & -v_NY_N & -v_N \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \\ h_{33} \end{bmatrix} = \mathbf{0} \tag{S3}$$

For simplicity, Equation (S3) is rewritten as $\mathbf{L} \cdot \mathbf{h} = \mathbf{0}$, where $\mathbf{L} \in \mathcal{R}^{2N \times 9}$ and $\mathbf{h} \in \mathcal{R}^{9 \times 1}$. The constraint condition $\|\mathbf{H}\| = 1$ is supplemented since $\mathbf{H} \leftrightarrow \gamma\mathbf{H}\,(\gamma \neq 0)$. Note that there are 8 degrees of freedom even though $\mathbf{H}$ contains 9 unknowns, thereby 4 pairs of feature matching points (i.e., 8 equations) are sufficient to find a closed-form solution.

Based on the least square method (LSM), the constraint condition $\mathbf{h}^T\mathbf{h} = 1$ is employed to prevent the generation of the invalid solution $\mathbf{h} = \mathbf{0}$. As a result, we perform the singular value decomposition (SVD) on matrix $\mathbf{L}$:

$$\mathbf{L} = \mathbf{U} \sum \mathbf{V} \tag{S4}$$

Here, $\mathbf{V}$ consists of eigenvectors of a real symmetric matrix $\mathbf{L}^T\mathbf{L}$. By conducting the eigenvalue decomposition on $\mathbf{L}^T\mathbf{L}$, the homography vector $\mathbf{h}$ is deduced as a unit eigenvector corresponding to the smallest eigenvalue. Later, 4 pairs of feature matching points are selected from the satellite picture, whose real-world coordinates and image coordinates are measured by fieldwork and software. Fig. S1(c) shows the frame images are calibrated for perspective distortion using homography vector $\mathbf{h}$, we then choose feature matching point 1 as the coordinate origin and establish a real-world coordinate system.

D. Trajectory Cleaning and Analysis

The collected video with pedestrian tracking information is sampled at a time step Δt of 1 s (25 FPS), and we define $\Gamma(t)$ as the frame image at time t. The bottom center coordinate of the bounding box is regarded as the position coordinate of pedestrian i in frame image $\Gamma(t)$, which is then converted to a real-world coordinate using the calibration of perspective distortion. We collect a total of 311,000 position coordinates and remove 11,164 of them containing false detections (e.g., non-pedestrian objects, out-of-boundary detections) by manually checking 86,400 frame images. From this, the trajectory of pedestrian i in continuous frame images is expressed as $[\mathbf{r}_i(t_i^s), \cdots, \mathbf{r}_i(t_i^e)]$, where t_i^s is the time when pedestrian i appears on the screen, and t_i^e is the time before he or she disappears from the screen.

Due to the limitation of perspective, however, the tracking effect is interfered by the mutual occlusion among pedestrians to a certain extent. This results in missing values in pedestrian trajectories, whereby it is required to clean the trajectory data for ensuring its availability and accuracy. It is assumed that the trajectory of pedestrian i is incomplete, which is written as $[\mathbf{r}_i(t_i^s), \cdots, \mathbf{r}_i(t_i^p), \mathbf{r}_i(t_i^{p+q\Delta t}), \cdots, \mathbf{r}_i(t_i^e)]$, where q is a positive integer and satisfies $q \geq 2$. Given that the corresponding velocities in this trajectory can also be easily calculated, we first consider interpolating the missing velocities by assuming a constant acceleration:

$$\mathbf{v}_i(t_i^{p+\delta\Delta t}) = \mathbf{v}_i(t_i^p) + \frac{\delta}{q}\left(\mathbf{v}_i(t_i^{p+q\Delta t}) - \mathbf{v}_i(t_i^p)\right) \tag{S5}$$

where $\delta \in \{1, \cdots, q-1\}$, and those missing positions in this trajectory are therefore generated as follows:

$$\mathbf{r}_i(t_i^{p+\delta\Delta t}) = \mathbf{r}_i(t_i^{p+(\delta-1)\Delta t}) + \mathbf{v}_i(t_i^{p+(\delta-1)\Delta t})\Delta t \tag{S6}$$

However, two issues are found to remain with these processed pedestrian trajectories. One is the trajectories with short durations, due to the rapid appearance and disappearance of pedestrians around border areas. The other is the trajectories with long durations, caused by staying for a long time or incorrectly matching at the confluence. To address the above issues, the trajectories whose durations satisfy $t_i^e - t_i^s \leq t_d^{\min}$ are deleted because they are trivial for subsequent analysis. For the trajectories whose durations tally with $t_i^e - t_i^s \geq t_d^{\max}$, part of which formed by prolonged stay are directly deleted, and others are assigned new IDs to correct the incorrectly matched bounding boxes. Note that $t_d^{\min}$ and $t_d^{\max}$ are set to 5 s and 60 s according to video observations. Finally, as shown in Fig. S1(d), 12,326 valid pedestrian trajectories are extracted based on the above series of steps.

The temporal series analysis of pedestrian trajectories is also performed to better understand the fundamental quantities (e.g., density, speed, and flow) of crowd movements [8]. The density $\rho(t) = N_t/A_S$ is defined as the number of pedestrians per unit area, where N_t denotes the number of pedestrians at time t. The speed is calculated using the general approach $v(t) = \sum_{i \in \Phi_S} \|\mathbf{v}_i(t)\|/N_t$, where $\mathbf{v}_i(t)$ is the velocity of pedestrian i at time t, and Φ_S is the set of all pedestrians in study area S. The flow is deduced based on the fluid-dynamic equation $Q(t) = \rho(t) \cdot v(t)$. Fig. S1(e) depicts the 24-hour temporal series of density, speed, and flow calculated based on the extracted trajectories. The periods during breakfast time $(06:40-07:20)$, lunch time $(11:50-12:50)$, dinner time $(17:00-18:00)$, and supper time $(22:00-22:40)$ are held as four concentrated phases of crowd emergence, while a small number of pedestrians can also be observed passing through the study area during free time in the video. The above facts demonstrate that the characteristics of crowd movements in this environment are complicated, which brings considerable difficulty to the automatic identification of subgroups.

REFERENCES

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[2] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2017.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[4] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, sep 2015.
[6] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.
[7] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Comput. Soc, 1997.
[8] H. Dong, M. Zhou, Q. Wang, X. Yang, and F.-Y. Wang, "State-of-the-art pedestrian and evacuation dynamics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1849–1866, may 2020.

Wenhan Wu received the B.S. degree in School of Automation from Central South University, Changsha, China, in 2019. He is currently pursuing the Ph.D. degree in control science and engineering with the Tsinghua University, Beijing, China. His current research interests include collective behavior, emergency evacuation and pedestrian group dynamics.

Wenfeng Yi received the B.S. degree in School of Astronautics from Beihang University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree in control science and engineering with the Tsinghua University, Beijing, China. His current research interests include collective behavior, emergency evacuation and pedestrian queueing dynamics.

Jinghai Li received the B.S. degree in automation and the M.S. degree in control science and engineering from Tianjin University, Tianjin, China, in 2009 and 2016, respectively, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2022. He is currently a Postdoctoral Researcher at the School of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, Beijing, China. His current research interests include crowd dynamics, control of robotic systems, and adaptive systems.

Maoyin Chen (Member, IEEE) received the B.S. degree in mathematics and the M.S. degree in control theory and control engineering from Qufu Normal University, Shandong, China, in 1997 and 2000, respectively, and the Ph.D. degree in control theory and control engineering from Shanghai Jiao Tong University, Shanghai, China, in 2003.

From 2003 to 2005, he was a Postdoctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China. From 2006 to 2008, he visited Potsdam University, Potsdam, Germany, as an Alexander von Humboldt Research Fellow. Since October 2008, he has been an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. He has authored and coauthored over 100 peer-reviewed international journal papers. He has won the first prize in natural science (2011, ranked first) and the second prize (2019, ranked first) of CAA. His research interests include fault prognosis and complex systems.

Xiaoping Zheng received the B.S. degree from the Chengdu University of TCM, Chengdu, China, in 1995, and the Ph.D. degree from Sichuan University, Chengdu, China, in 2003.

From 2004 to 2006, he was a Postdoctoral Researcher in School of Management, Fudan University, Shanghai, China. From 2006 to 2013, he was a Professor with the Institute of Safety Management, Beijing University of Chemical Technology, Beijing, China. He is currently a Professor with the Department of Automation, Tsinghua University, Beijing, China. Prof. Zheng was a 973 Chief Scientist in 2011, a recipient of the National Science Fund for Distinguished Young Scholars in 2012, and a Distinguished Professor of the Chang Jiang Scholars Program in 2021. His current research interests include large-scale crowd evacuation, evolutionary game theory and Terahertz technology.