Collective Dynamics

Benchmarking High-Fidelity Pedestrian Tracking Systems for Research, Real-Time Monitoring and Crowd Control

Caspar A.S. Pouw[1,3] · Joris Willems[2] · Frank van Schadewijk[3] · Jasmin Thurau[4] · Federico Toschi[1,5] · Alessandro Corbetta[1]

[1] Department of Applied Physics, Eindhoven University of Technology, Eindhoven, The Netherlands
E-mail: c.a.s.pouw@tue.nl, f.toschi@tue.nl, a.corbetta@tue.nl
[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands
E-mail: j.e.m.willems@student.tue.nl
[3] ProRail Stations, Utrecht, The Netherlands
E-mail: cas.pouw@prorail.nl, frank.vanschadewijk@prorail.nl
[4] SBB AG, Bern, Switzerland
E-mail: jasmin.thurau@sbb.ch
[5] CNR-IAC, Rome, Italy

Received: 25 August 2021 / Last revision received: 14 November 2021 / Accepted: 22 November 2021
DOI: 10.17815/CD.2021.134

Abstract High-fidelity pedestrian tracking in real-life conditions has been an important tool in fundamental crowd dynamics research allowing to quantify statistics of relevant observables including walking velocities, mutual distances and body orientations. As this technology advances, it is becoming increasingly useful also in society. In fact, continued urbanization is overwhelming existing pedestrian infrastructures such as transportation hubs and stations, generating an urgent need for real-time highly-accurate usage data, aiming both at flow monitoring and dynamics understanding. To successfully employ pedestrian tracking techniques in research and technology, it is crucial to validate and benchmark them for accuracy. This is not only necessary to guarantee data quality, but also to identify systematic errors. Currently, there is no established policy in this context.

In this contribution, we present and discuss a benchmark suite, towards an open standard in the community, for privacy-respectful pedestrian tracking techniques. The suite is technology-independent and it is applicable to academic and commercial pedestrian tracking systems, operating both in lab environments and real-life conditions. The benchmark suite consists of 5 tests addressing specific aspects of pedestrian tracking quality,

including accurate line-based crowd flux estimation, global density estimation, individual position detection and trajectory accuracy. The output of the tests are quality factors expressed as single numbers. We provide the benchmark results for two tracking systems, both operating in real-life, one commercial, and the other based on overhead depth-maps developed at TU Eindhoven, within the Crowdflow topical group. We discuss the results on the basis of the quality factors and report on the typical sensor and algorithmic performance. This enables us to highlight the current state-of-the-art, its limitations and provide installation recommendations, with specific attention to multi-sensor setups and data stitching.

Keywords High-fidelity pedestrian tracking · sensor benchmarking · crowd monitoring · real-life pedestrian measurements · industrial and societal applications

1. Introduction

Growing population and continued urbanization puts urban infrastructures at large stress. Moreover, over the next 10 years in densely populated European countries public transport facilities such as, e.g. train or metro stations, expect a passenger growth as high as 40% [1]. Potentially dangerous crowd-capacity issues – possibly in combination with distancing requirements – increase by the day, and demand substantial crowd management efforts. To unlock sustainable and scalable crowd management, maximizing comfort and safety, real-time, high-accuracy, anonymous individual pedestrian tracking is a must. This enables reliable usage monitoring and performance profiling, and, on a broader perspective, the possibility to develop a fundamental understanding of the motion of crowd flows. Pedestrian dynamics researchers who, since the beginning of the millenium, mostly relied on controlled laboratory experiments for data acquisition (see e.g. [2–8]), can now also acquire fundamental knowledge in real-life environments collecting large-scale statistics of observables such as walking velocities, mutual distances, body orientations and group structures [9–14].

Optic-based tracking, leveraging on visual-like signals, is the most widespread technology when (sub-)centimeter pedestrian positioning resolution is required. Raw signals are generally acquired via arrays of color cameras (as CCTV) [2,5], stereo cameras [3,14,15] or infrared depth-cameras [9, 16, 17]. Specifically, the last two technologies, hinged on three-dimensional imaging, allow higher accuracy and will be considered in this paper. After a calibration stage in which, among others, pixel-coordinates are matched to spatial coordinates, raw signals are post-processed to yield pedestrian positions and trajectories. Highly accurate, optical methods are generally limited in range by the visual cone of the individual sensors. Note that, at the price of a substantial accuracy loss, Bluetooth-based [18, 19] or Wi-Fi-based [20, 21] tracking enable larger spatial coverage per sensor.

Optic-based tracking techniques become rapidly more affordable over time, this makes high-accuracy pedestrian tracking accessible to a wide variety of users beyond academic research. Over the last few years, for instance, managers of public transport facilities have adopted pedestrian tracking technologies to gain valuable insights in the flow dynamics through their facilities (see e.g. [14, 22–24]).

To successfully employ pedestrian tracking techniques in research and technology, it is crucial to measure the level of performance and asses the quality of the output data. This means establishing a benchmark as well as measuring the performance of the current state-of-the-art. This is not only necessary to guarantee data quality, but also to identify systematic errors, such as erroneous object-person recognition, and misaligned stitching. To the best of our knowledge, there is no established tool in this context.

In this contribution we propose a benchmark that consists of a minimal set of five tests to quantify pedestrian detection accuracy and time-tracking reliability. The tests have been iteratively improved over the past 3 years [22] and follow from a joint effort between academic research and large-scale public facility managers. The full benchmark, designed to take limited time and resources, can be executed in less than 2 hours with as few as 12 participants. During the test participants are asked to walk close to each other. Due to short distances between first neighbors we are able to obtain instantaneous high local densities despite low number of participants. The five distinct tests span from macroscopic to increasingly microscopic observables of pedestrian dynamics. Considering finer and finer aspects of pedestrian dynamics, the tests get increasingly challenging from a technological perspective. The tests output quality factors expressed as single numbers. The first two tests probe large-scale observables by gauging reliability in estimating: 1. crowd fluxes (ped/min), and 2. local densities (ped/m^2). Note that these are averaged quantities, respectively over a time interval or over a surface, and therefore benefit from error compensation, i.e. false negatives could counterbalance false positives. In test 3, we focus on the significantly more challenging task of instantaneous individual localization. In tests 4 and 5, we consider Lagrangian time-tracking proficiency over full multi-sensor measurement domains.

We present the benchmark results for two pedestrian tracking setups operating in real-life, one developed in-house at TU Eindhoven, and the other one commercially available. This aims at reporting on the current state-of-the-art and providing a reference for new pedestrian tracking setups.

This paper is structured as follows: In Sec. 2 we discuss the essentials of optic-based pedestrian tracking. This is followed, in Sec. 3, by a description of the two experimental setups we employ to validate our benchmark. In Sec. 4, we introduce the benchmark suite and detail individual tests and their rationale. In Sec. 5, we report the benchmark performance of our experimental setups. The discussion in Sec. 6 concludes the paper.

2. Essentials of 3D optic-based pedestrian tracking

In 3D optic-based tracking, visible-light or infrared 3D imaging data, acquired by camera-like sensors, are processed to localize pedestrians in space and track them over time. Three-dimensional imaging, richer in information than flat 2D pictures, substantially simplifies and allows high accuracy in the tasks connected to tracking. At the core it is the estimation of a depth-map of the scene, that encodes the position of each pixel in the three-dimensional space [12, 16, 17]. This can be achieved via stereoscopic vision, scattered infrared illumination e.g. [25], or time-of-flight sensors. For each pedestrian and

each 3D-frame acquired, the tracking process yields quadruplets (x, y, t, id) where x and y are spatial coordinates (a z-coordinate can be added if needed), t is the frame acquisition time, and id is an identifier unique to each pedestrian. This enables us to define the set of recorded trajectories $\mathscr{T} = \{\gamma_{id}\}$, with γ_{id} the trajectory of the pedestrian with identifier id.

The tracking process generally happens in two stages: localization and Lagrangian time-tracking. In the localization stage, each frame is processed independently to single out each pedestrian and estimate their position. To this purpose image processing and/or machine learning models are used [13, 16, 17, 26]. Lagrangian time-tracking assigns an id to each detection on the basis of continuity arguments. These two stages can also happen simultaneously, e.g. using optical-flow like techniques [15].

Cameras or infrared sensors are generally mounted in overhead position, aimed perpendicularly to the floor to reduce mutual pedestrian occlusions. The scope of their view-cone depends on, and in general, is limited to, the mounting height (typically the height of the ceiling). Grids of sensors with partially overlapping view-cones can be combined to enlarge the measurement domain [27]. During the installation, a calibration step which establishes a global coordinate system across all the sensors is generally performed. Additionally, background subtraction which removes stationary objects (e.g. benches) from the image can be used to simplify the localization phase and increase its accuracy.

Despite its growing adoption, optic-based pedestrian tracking still features numerous open technical challenges. First, localization algorithms can fail due to poor image quality, e.g. caused by impurities on the camera lenses, excessive or insufficient illumination, or interference in the infrared spectrum (caused by direct sun exposure). Additionally, objects in the scene such as luggage or bicycles can yield false positive detections. Instead, possible causes for false negatives are (partial) occlusions by other pedestrians or infrastructural objects, e.g. trusses, signage, and lighting. Especially at the boundaries of each sensors view-cone the acquired image can be distorted which also lowers the localization quality. False negatives in the localization process yield gaps over which the time-tracking algorithm is unable to return continuous trajectories. As a consequence, two or more "interrupted" trajectory pieces with distinct ids are returned.

Combining information from multiple sensors presents also challenges connected to insufficient overlap or misalignment between the sensor view-cones (cf., e.g., the view-cone stitching algorithm in [12]). These are typical causes for "interrupted" trajectories.

Our benchmark contains tests that are designed to recognize these bottlenecks, probing efficiently localization accuracy, geometric conformality, and the absence of tracking artifacts such as misassigned ids.

3. Tracking technologies and experimental setups

We discuss our benchmark considering two pedestrian tracking setups briefly described below. The systems, which operate anonymously and in real-life conditions, leverage on different optic-based tracking approaches. The first, developed in house at TU Eindhoven, within the Crowdflow topical group, is based on depth reconstructed via scattered infrared

illumination [25] and, the second, is based on depth reconstructed via stereoscopic vision. In the following we shall refer to these systems, respectively, as "TU/e setup" and "commercial setup". We give a description of the pedestrian tracking systems in the paragraphs below. Note that in both setups we did not apply any further processing of the raw images nor post-processed the obtained trajectories. The performance of these tracking systems can be enhanced by setup-specific manual operations e.g. smoothing or restitching of the individual trajectories. This choice was deliberately made to focus on the bare, baseline, tracking accuracy.

TU/e setup The tracking system developed in house at TU Eindhoven leverages on overhead depth-map images, which represent the distance between pixels and the sensor plane (colorized in shades of gray in the example of Fig. 1c). Localization occurs via depth clustering (as in [12]), and time-tracking uses the Trackpy Python library [28]. The same approach has been successfully used, e.g., in stations, streets, and museums [9, 12, 13, 26]. The specific setup considered consists of a grid of 3×4 Microsoft Kinect$^{\text{TM}}$ [25] depth sensors. The sensors are attached to the ceiling of a large public area within the University campus in Eindhoven, the Netherlands, at a height of about 4.5 m (see Fig. 1a, b). A staircase (visible in Fig. 1b) is located next to, but not part of, the measurement setup. The grid records depth images (Fig. 1c) over an area of $S = 150$ m^2 with $f = 30$ frames per second.

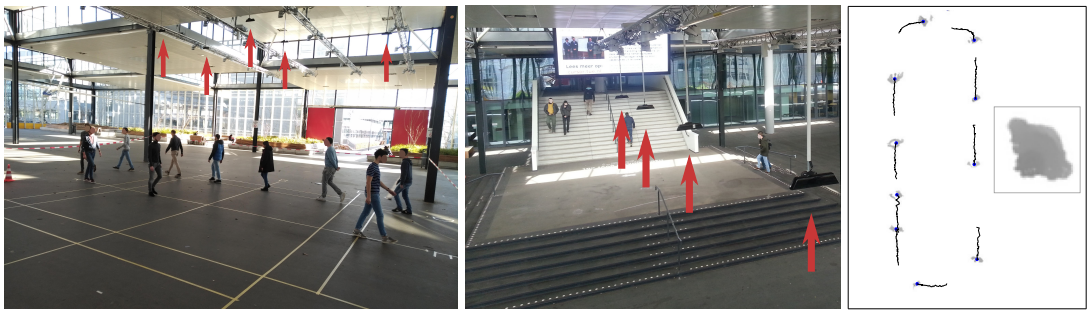
Commercial setup The commercial system anonymously tracks pedestrian movements using 3D stereoscopic images. The system consists of 3 commercial pedestrian tracking sensors (Xovis$^{\text{TM}}$), used in e.g. train stations [14, 22–24], to monitor complex crowd flows. Every sensor records images at $f = 10$ frames per second and processes the stereo images in real-time only storing pedestrian locations as x, y coordinate pairs. This system is installed at real-life operational train station Breukelen, the Netherlands. The sensors are mounted to the ceiling of a platform covering an area of $4\,\text{m} \times 12.4\,\text{m} \approx 50\,\text{m}^2$. We report an overview of the platform in Fig. 2.

Note that (most) commercial systems only return trajectories, whereas in custom setups one can retain also the raw data which can be used to further improve the algorithms and/or extracting additional features such as body orientations.

4. Pedestrian tracking benchmark

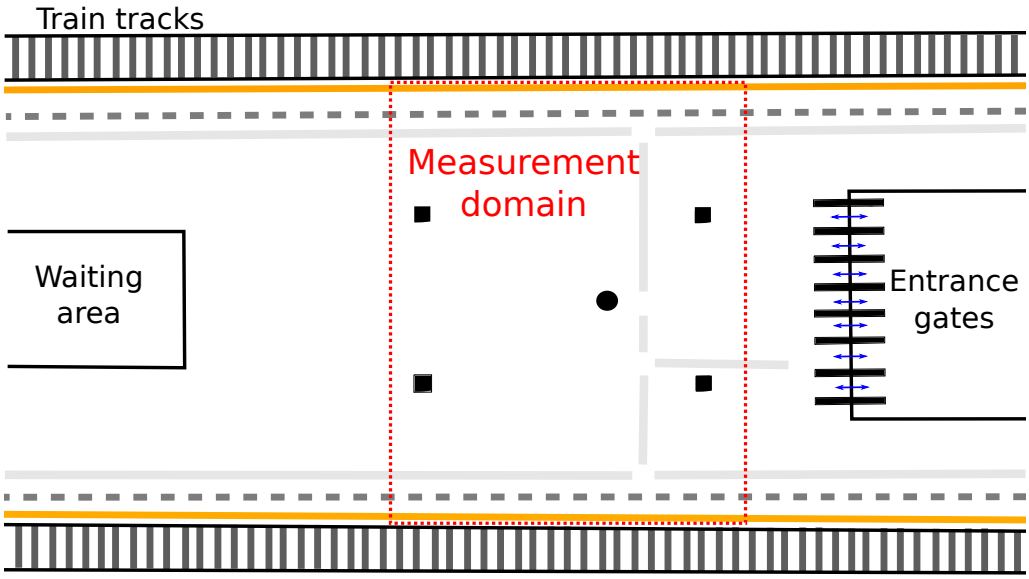
We describe here the five tests comprised in our benchmark, which is also resumed and schematized in Tab. 1 encompassing an illustration and key features for the individual tests. We additionally present advises on test duration, ΔT, and number of participants, N, using a gray cell background.

Test 1: Line-based crowd flux estimation. We probe the accuracy in estimating crowd fluxes, as count of pedestrians crossing a line in a predefined time window, $t_0 < t < t_1$.



(a) (b) (c)

Figure 1 TU/e measurement setup based on overhead depth sensors at Eindhoven University of Technology, The Netherlands, during test 3. (a) Picture taken during test 3 of the benchmark. At the top of the image part of the 3×4 sensor grid is indicated with red arrows. We guide participants during this test using taped markings on the floor. (b) Detail of one of the three arrays of 4 sensors, taken at sensor height. The sensors are attached to a truss near the ceiling pointing downward, perpendicularly to the floor (c) Overhead depth image captured by the sensor grid already encompassing merging and perspective correction [29]. The gray color in the depth-map represents the distance to the camera plane. Bright shades are far from the sensor and darker colors are closer to the sensor. In the depth-map we can distinguish silhouettes of pedestrians (as seen from above) where the shoulders are a lighter tint of gray and the head is slightly darker. The inset reports the silhouette of a single pedestrian. The trajectories of the pedestrians are super-imposed and show the walking direction.



(a)

Figure 2 Experimental setup for the commercial tracking system at train station Breukelen, The Netherlands, visualized with a schematic floorplan. The train tracks are shown in the top and bottom of the image. The dimensions of the measurement domain, with size 4 m $\times$ 12.4 m $= 50$ m^2, is highlighted with a red dashed rectangle.

We compare counts automatically estimated, $N(t_0,t_1)$, with the ground-truth, $N_{real}(t_0,t_1)$, manually evaluated. As a final score we retain the following indicator

$$A^{(1)} = \left(1 - \frac{|N(t_0,t_1) - N_{real}(t_0,t_1)|}{N_{real}(t_0,t_1)}\right) \cdot 100, \tag{1}$$

which equals 100 in case of a correct count estimate and is lower otherwise. Commercial systems generally include internal algorithms to compute line crossings, whereas in the TU/e system we employ the algorithm described in App. A. To reach a challenging flux of $J_A \approx 100$ ped/min, we ask participants to follow an elliptical path with a small circumference ($C \approx 10$ m). The elliptical path is crossed twice by the straight line across which the flux estimation occurs. The test takes $\Delta T \approx 5$ minutes to complete.

Test 2: Global density estimation. We target the accuracy in estimating pedestrian densities. We consider the number of people $N_1, N_2, N_3, \ldots$ moving freely within virtual regions $S_1, S_2, S_3, \ldots$ determined by the tracking systems, and compare it with the ground-truth. For simplicity, we keep the number of pedestrians in each virtual region constant, $N_{i,real}$, and consider the following time-averaged relative error as the score

$$A_i^{(2)} = \left(1 - \frac{1}{T}\int_0^T \frac{|N_i(t) - N_{i,real}|}{N_{i,real}}\, dt\right) \cdot 100. \tag{2}$$

With few participants we only target low average crowd densities in this test. However, the localization task is very challenging due to short distances between first neighbors that yield instantaneously local high densities. Additionally, (stationary) objects can be added to the test area to validate the ability to differentiate between objects and people.

Test 3: Individual position detection. We target, in line with [22], the capability of accurately determining individual positions. To bypass the need to manually establish a ground-truth for point-wise comparisons, we ask participants to walk following simple geometric patterns, specifically, a grid of straight lines (cf. markings in Fig. 1a). We score how closely the measured trajectories agree, as an ensemble, with the geometric pattern, i.e. they form thin and straight bands.

Operationally, for each collected trajectory, we isolate the portions that follow single grid lines. For each grid line, we obtain a set of trajectory pieces of which we consider averages. We either retain the best fitting straight line (linear regression) or we find a piece-wise average in bins of $D_{bin} = 5$ cm (local regression). Naming, without loss of generality, these fitting curves, respectively, $y_{lin} = y_{lin}(x)$ and $y_{loc} = y_{loc}(x)$, we quantify the following (the segments naming is as in Tab. 1 third entry):

- spread of trajectories along each grid line, that should be comparable to the individual pedestrians body sway amplitude (about 5 cm [30]). We consider for each x-location parametrizing a line the quantity $z_{lin}(x) = y_{lin}(x) - y(x)$ where $y(x)$ is a

generic measurement of coordinate y at position x. Note that $z(x)$ is approximately zero-centered at each x. Our benchmark quantifies

$$\sigma_{lin}^{(3)} = \underset{\substack{\text{measurements} \\ \text{at all } x}}{\text{std.dev.}}(z_{lin}(x)), \tag{3}$$

likewise it holds for $\sigma_{loc}^{(3)}$;

- distance between linear fits over parallel lines, that should be constant. We score them with slopes, such as the following, for the case of segments AK, BL (for the other segments the formula works similarly)

$$D^{(3)} = 100 - \left(\frac{|DE| - |HI|}{|DH|}\right) \cdot 100; \tag{4}$$

- angles between linear fits over perpendicular lines, that should be 90°. For the angle $\angle DHI$ we score this as follows (generalization for the other angles is not reported)

$$L^{(3)} = 100 - \left(\frac{|\angle DHI - 90^\circ|}{90^\circ}\right) \cdot 100. \tag{5}$$

Test 4: Trajectory accuracy in controlled environment We target the capability of tracking pedestrians for extended time periods and along complex trajectories. We divide the measurement domain in regions $S_1, S_2, S_3, \ldots$, and ask each participant, id, from a set of N_{real} pedestrians, to stand in a region S_o^{id} and walk to an assigned destination S_d^{id}, following an irregular path of choice taking roughly $\Delta T \approx 30$ s. Origin and destination regions are assigned exclusively to a single pedestrian. With few participants, e.g. $N_{real} \approx 12$, the average density is low during this test. However, instantaneously we have extremely high densities due to short distances between first neighbors, this makes tracking very challenging. A recorded trajectory is considered correct when its origin and destination respectively lie within the boundaries of the assigned origin-destination pair (S_o^{id}, S_d^{id}). The final score is the percentage of correct trajectories. To increase the difficulty, (stationary) pedestrians standing outside all the regions $S_1, S_2, S_3, \ldots$, can be added to the measurement domain

$$A^{(4)} = \left(\frac{N_{corr}}{N_{real}}\right) \cdot 100. \tag{6}$$

Test 5: Trajectory accuracy in real-life environment Finally, we test in a real-life environment the capability of tracking pedestrians without interruptions. We define an inner region, S_{in}, in which no trajectory can physically start or terminate. Each trajectory recorded in a time window $\Delta T \geq 1$ day, that enters the domain S_{in}, is classified according to its quality:

1. **correct**: neither the trajectories initial nor final point lay inside region S_{in};
2. **faulty termination**: the trajectory terminates inside region S_{in};

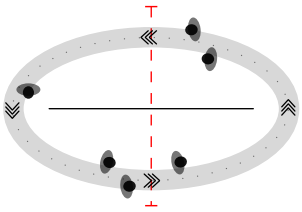
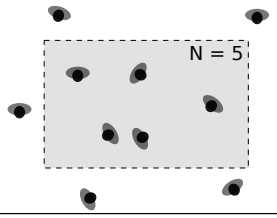
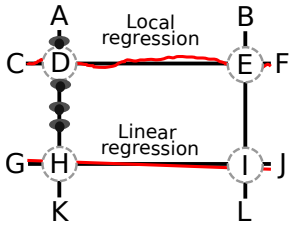
3. **faulty origin**: the trajectory originates inside region S_{in}.

We report the percentage of trajectories correctly tracked $A(5)$, approximating the total number of trajectories as the correct trajectories plus interrupted trajectories. In formulas this reads

$$A^{(5)} = \left(\frac{\text{correct}}{\text{correct} + \text{interrupted}} \right) \cdot 100, \tag{7}$$

where interrupted trajectories are interrupted paths consisting of two or more trajectory pieces. Because one trajectory piece must enter domain S_{in} and another must leave this domain we can approximate interrupted trajectories as

$$\text{interrupted} = \frac{\text{faulty termination} + \text{faulty origin}}{2}. \tag{8}$$

Features		Illustration	Description
Test	Crowd-flux estimation		Participants walk along an elliptical path, thereby crossing a virtual line (in red). The test reports for every minute the error between the sensor estimated and the ground-truth crowd-flux across the (red) virtual line.
Metric	$A^{(1)}$ (Eq. 1)		
ΔT	5×1 min		
N	12		
Test	Density estimation		The number of participants inside a predefined area is kept constant (e.g. N = 8). The test reports the relative error between the estimated, $N(t)$, and ground-truth number of pedestrians, N_{real}, inside the area.
Metric	$\varepsilon^{(2)}$ (Eq. 2)		
ΔT	4×5 min		
N	12		
Test	Individual position detection		Participants walk in a row in a straight line. The test reports the standard deviation in the distance between the recorded data point and the local and linear regressions. Additionally, we report the angles and distances between the linear regressions to identify distortions in the measurement setup.
Metrics	$\sigma^{(3)}$ (Eq. 3) $D^{(3)}$ (Eq. 4) $L^{(3)}$ (Eq. 5)		
ΔT	2×5 min		
N	12		

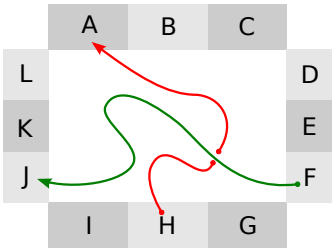
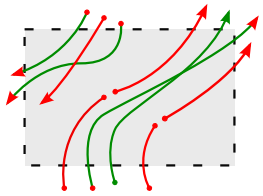
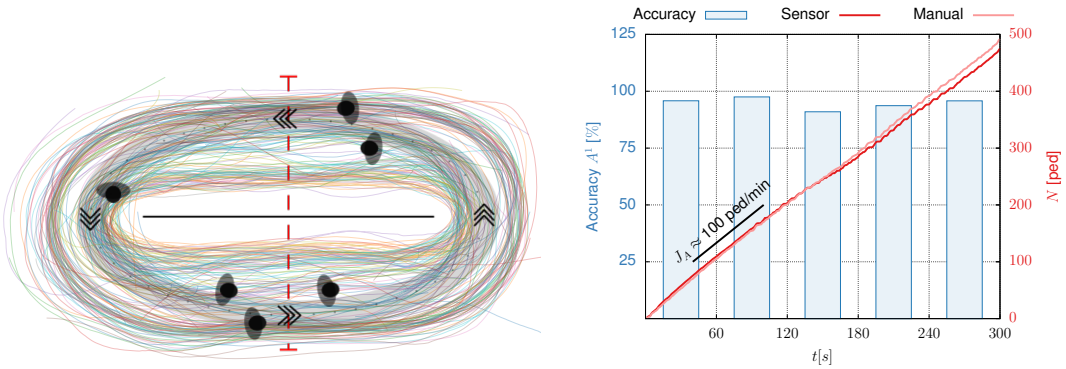
Test	Trajectory accuracy		Participants walk in an irregular path from a predefined origin to a predefined destination, passing in close proximity from each other. The test reports the percentage of trajectories that is accurately tracked from origin to destination without interruption.
Metric	$A^{(4)}$ (Eq. 6)		
ΔT	4×1 min		
N	12		
Test	Trajectory accuracy		We define a domain, well within the sensor range, where no trajectory should originate or terminate. The test reports in real-life conditions the percentage of trajectories that is continuously tracked through that domain.
Metric	$A^{(5)}$ (Eq. 7)		
Env	Real-life		
ΔT	≥ 1 day		

Table 1 Synthetic description of the benchmark tests including key features and an illustration. The metric used to score the test which we supplemented with advised quantities for the test duration, ΔT, and number of participants, N, is included. Cells with advised quantities have a gray background.

5. Benchmark results on the considered systems

In this section we report and elaborate on the benchmark results for the pedestrian tracking systems introduced in Sec. 3. We iteratively improved each test by trying different methods and variants. This optimization process, in combination with the ever-changing nature of the real-life testing environments, caused, minor, non-essential, differences between the two setups. For each test, we report synthetic results plus illustrations taken from either setup.

Test 1: Line-based crowd flux estimation. In Fig. 3a we report a sample of captured trajectories and the results, in graph form, for the TU/e measurement setup. Specifically, in Fig. 3b we report the minute-by-minute estimation accuracy (blue bars), the cumulative pedestrian count (red line), and the average crowd flux (black slope). Synthetic results for both setups are in Tab. 2, which includes, in time windows $t_0 < t < t_1$ of 1 minute, the ground-truth pedestrians count, $N_{real}(t_0,t_1)$, the crowd flux estimation error, $N(t_0,t_1) - N_{real}(t_0,t_1)$, and the estimation accuracy, $A^{(1)}$ (Eq. 1). For the TU/e system, we employ $N = 12$ participants walking an elliptical path with circumference $C \approx 10$ m, whereas for the commercial setup $N = 11$ participants walk an ellipse with circumference $C \approx 22$ m. As a consequence of the different path lengths, C, the commercial setup is tested with an average crowd flux of $J_A \approx 30$ ped/min, whereas the TU/e measurement setup is tested with a more challenging $J_A \approx 100$ ped/min. Results of the TU/e and commercial setups are $A^{(1)} = 95\%$ and $A^{(1)} = 100\%$ respectively.

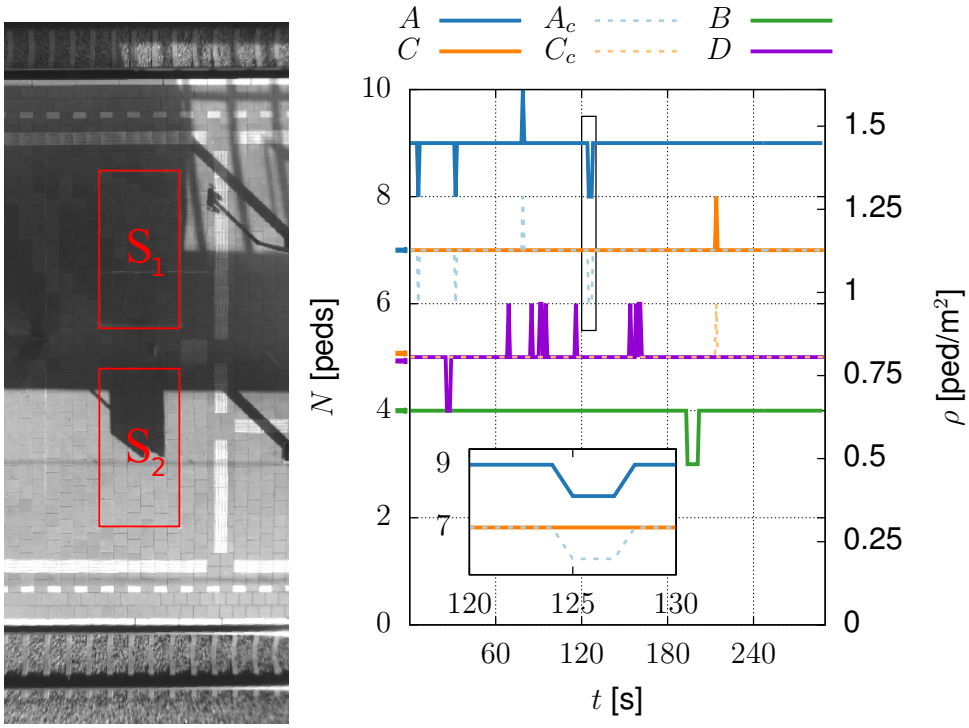


(a) (b)

Figure 3 (a) Trajectories captured during test 1 by the TU/e setup, superimposed to the schematic illustration from Tab. 1. (b) Results of test 1 for the TU/e setup. Blue bars report the accuracy, $A^{(1)}$, on a minute-by-minute basis, red lines indicate measured (dark) and ground-truth (light) pedestrian count, $N(t)$, and a black line indicates the slope i.e. the average crowd flux, $J_A \approx 100$ ped/min.

Table 2 Synthetic results of test 1 for both pedestrian tracking setups. We report, on a minute-by-minute basis, the ground-truth number of pedestrians, N_{real}, the error in the count, $N - N_{real}$, and the estimation accuracy, $A^{(1)}$. While testing the TU/e setup, the participants maintained an average mutual distance of $d_{avg} \approx 0.8$ m yielding a crowd flux of $J_A = 100$ ped/min, whereas in the commercial case a far less challenging crowd flux of $J_A = 30$ ped/min was maintained due to the larger average distance between first neighbors $d_{avg} \approx 2$ m. Therefore, the test of the commercial system can be considered relatively less difficult.

Minute	**TU/e** $J_A \approx 100$ ped/min $d_{avg} \approx 0.8$ m			**Commercial** $J_A \approx 30$ ped/min $d_{avg} \approx 2$ m		
	N_{real}	$N - N_{real}$	$A^{(1)}$	N_{real}	$N - N_{real}$	$A^{(1)}$
1	105	4	96%	17	0	100%
2	96	2	98%	38	0	100%
3	90	8	91%	37	0	100%
4	98	6	94%	29	0	100%
5	103	4	96%	NA	NA	NA
Total	492	24	95%	121	0	100%



(a) (b)

Figure 4 (a) Measurement domain for the commercial setup. Red domains indicate regions S_1 and S_2 both with an area 6.2 m^2. (b) Results of test 2 for the commercial setup. The figure reports for all repetitions, $(A-D)$, the measured pedestrian count, $N(t)$. A color matching mark on the left y-axis indicates the ground-truth pedestrian count N_{real} for each repetition. Additionally, we report for A and C a systematic error corrected pedestrian count denoted A_c and C_c respectively. The left y-axis indicates pedestrian count whereas the right y-axis shows the corresponding crowd density ρ inside the 6.2 m^2 region. The inset shows an enlarged view of the graph part inside the black box.

Test 2: Global density estimation. We report, for the commercial setup, the test layout in Fig. 4a and the test results in Fig. 4b. For the commercial setup we define two regions S_1, S_2 (cf. Fig. 4a), both with an area of 6.2 m^2. In the case of the TU/e setup we employed only one larger region $S = 150$ m^2. To improve test reliability, we perform multiple runs for each setup, thereby realizing four density estimations each $A-D$, see Tab. 3. In the table we report also the ground-truth region occupation, N_{real}, the density, $\rho = \frac{N_{real}}{S}$, and the estimation accuracy $A^{(2)}$ (Eq. 2). The commercial setup shows a systematic error in the density estimation of region S_1, most likely related to false positive detection of 2 stationary objects. Therefore, in Tab. 3, we include an additional column containing a corrected estimate $A_c^{(2)}$, for which the systematic error is removed by reducing the pedestrian count, $N(t)$, by 2.

Among all four tests the commercial setup sustained an average density of $\rho = 0.85$ ped/m^2, whereas the density across the TU/e measurement domain is a factor 10 lower. Throughout run D, the infrared sensor of the TU/e setup is overexposed by

excessive sunlight. Reduced image quality causing false negatives results, for this run, in a lower density estimation accuracy.

Test 3: Individual position detection. We report in Fig. 5a, for the commercial setup, the recorded trajectories during test 3. We fit a straight line (linear regression) for every isolated set of trajectory pieces. The regressions accurately reconstruct the geometric structure that is followed by the participants i.e. closely resemble a grid of straight lines. We added the angles and distances between the grid lines to emphasize the correspondence. Fig. 5b provides the isolated set of trajectory pieces belonging to grid line BE. We fit a local (blue) and linear (red) regression through the trajectory pieces. Additionally, we report the spread along the grid line with histograms for $z_{lin}(x)$ (top) and z_{loc} (bottom) annotated with test scores $\sigma^{(3)}_{lin}$ and $\sigma^{(3)}_{loc}$. We refer to Tab. 4 for the standard deviations, $\sigma^{(3)}_{lin}$ and $\sigma^{(3)}_{loc}$, for both setups. All standard deviations are the same order as typical body sway amplitude i.e. $\sigma^{(3)} = 5$ cm. The test is relatively more challenging for the TU/e setup which needs more sensors for the large measurement domain. In Tab. 5 we report the correspondence to the grid geometry in terms of the distance between trajectories on parallel grid lines $D^{(3)}$, and the angle between trajectories on perpendicular grid lines $L^{(4)}$. The grid geometry is reconstructed with high accuracy as the angles and the mutual distances in the recorded trajectories agree up to 99% with the original grid structure.

Test 4: Trajectory accuracy in controlled environment. In Fig. 6, we report side-by-side the trajectories recorded by the TU/e measurement setup during the three runs of test 4. Correct trajectories have a green and faulty trajectories a red color. Tab. 6 reports the test results for both setups indicating, for each run, the number of trajectories, N_{real}, the number of correct trajectories, N_{corr}, and the trajectory accuracy, $A^{(4)}$. The participants experience high instantaneous densities with almost body-to-body contact. Minimum mutual distances in the order 20 cm are recorded several times over the lifespan of their trajectories. Some tests of the commercial setup also contained additional stationary pedestrians to increase the global density, this is represented in the table with an extra column N_{obj}. The TU/e setup records, over an area of $S = 150$ m^2, 22 correct from a total of 30 trajectories whereas the commercial setup, on a much smaller area $S = 50$ m^2, scores 49 out of 64 trajectories. Both setups report a trajectory accuracy in the order of $A^{(4)} \approx 75\%$ (Eq. 6). This shows that in conditions with highly entangled trajectories tracking procedures can be imperfect, as only 75% of the trajectories are captured properly.

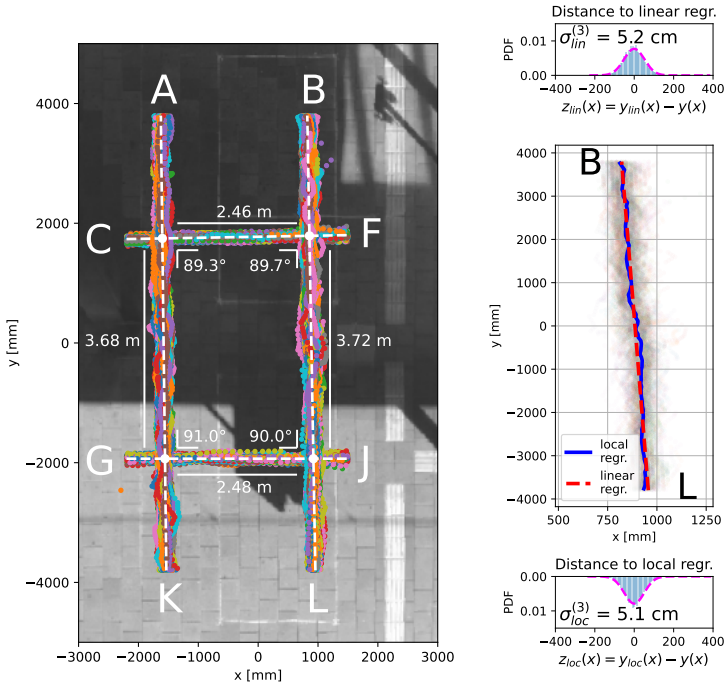
Test 5: Trajectory accuracy in real-life environment. In Fig. 7 we report all trajectories, recorded during $\Delta T = 1$ day, partitioned in subsets: correct, faulty termination, and faulty origin (cf. Sec. 4) using an inner domain $S_{in} \approx 38$ m^2. The first row of figures reports, per subset, the raw trajectories and the second row reports all the origins (red) and destinations (blue) of trajectories in the corresponding subsets. The percentage accurately tracked trajectories, $A^{(5)}$, is determined to be 79%, which is in the same order as the trajectory accuracy test in controlled conditions. This shows that under normal operational conditions 79% of the trajectory recordings is interrupted and broken into smaller pieces.

Table 3 Synthetic results of test 2 for both tracking setups. We report for the repetitions $A-D$ the ground-truth number of pedestrians, N_{real}, the estimated density, ρ and the estimation error, $\varepsilon^{(2)}$. The TU/e setup estimates density over an area of $S = 150\text{m}^2$ whereas the commercial setup considers much smaller regions of $S = 6.2\text{m}^2$. Due to the lower crowd density the test of the TU/e setup can be considered relatively less difficult. For the commercial setup we report additionally the estimation error after correction for systematic errors $A_c^{(2)}$.

Test	**TU/e** $S = 150\ \text{m}^2 \quad \Delta T \approx 300\ \text{s}$			**Commercial** $S = 6.2\ \text{m}^2 \quad \Delta T = 300\ \text{s}$			
	N_{real}	ρ [ped/m^2]	$A^{(2)}$	N_{real}	ρ [ped/m^2]	$A^{(2)}$	$A_c^{(2)}$
A	10	0.07	99%	7	1.13	71%	100%
B	8	0.05	99%	4	0.65	99%	99%
C	11	0.07	98%	5	0.81	60%	100%
D	12	0.08	94%	5	0.81	99%	99%
Total	41	0.07	97%	21	0.85	83%	100%

Table 4 Synthetic results of test 3 for both tracking setups. The table reports for each grid line, the number of overhead sensors, N_{sens}, and the standard deviation with respect to the local, σ_{loc} and to the linear, σ_{lin}, regression.

Test	**TU/e**			**Commercial**		
	N_{sens}	$\sigma_{loc}^{(3)}$ [cm]	$\sigma_{lin}^{(3)}$ [cm]	N_{sens}	σ_{loc} [cm]	σ_{lin} [cm]
AK	4	6.0	8.7	3	5.0	5.1
BL	4	4.9	5.5	3	5.1	5.2
CF	3	3.0	5.6	1	3.8	3.8
GJ	3	5.8	8.5	1	3.5	3.5
Average	3.5	4.9	7.1	2	4.4	4.4



(a) (b)

Figure 5 (a) Recorded trajectories by the commercial setup during test 3. The figure reports a linear fit for every set of trajectory pieces. The regressions reconstruct the geometric structure followed by the participants. Additionally, we report the distances between parallel fits and the angles between perpendicular fits. (b) Isolated trajectory pieces for grid line BL. We fitted a linear (red) and local (blue) regression through the trajectory pieces. Additionally, we report the histograms of $z_{lin}(x)$ (top) and $z_{loc}(x)$ bottom including a Gaussian fit (pink).

Table 5 Synthetic results of test 3 for both setups. The table reports how accurate the recorded trajectories agree with the grid geometry. In particular, we report the accuracy in reconstructing parallel grid lines, $D^{(3)}$, and the accuracy in reconstructing perpendicular grid lines, $L^{(3)}$.

Test	**TU/e**	**Commercial**
$D^{(3)}$	98.8 %	98.8 %
$L^{(3)}$	98.9 %	99.4 %

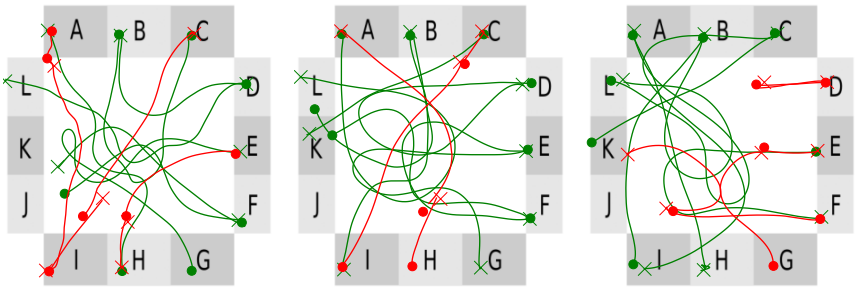


Figure 6 Trajectories captured by the TU/e setup during test 4. Green lines indicate correct trajectories whereas red lines indicate faulty trajectories. Trajectory origins and destinations are indicated with spheres and crosses respectively.

Table 6 Synthetic results of test 4 for both setups. We report for runs $A-D$ the number of participants, N_{real}, the number of correctly tracked trajectories, N_{cor}, and the test accuracy, $A^{(4)}$. Additionally we report for the commercial setup the number of (stationary) objects, N_{obj}.

Test	**TU/e** $S = 150$ m^2 $d_{min} \approx 20$ cm			**Commercial** $S = 50$ m^2 $d_{min} \approx 20$ cm			
	N_{real}	N_{cor}	$A^{(4)}$	N_{real}	N_{obj}	N_{cor}	$A^{(4)}$
A	10	7	70%	16	0	12	75%
B	10	8	80%	16	0	12	75%
C	10	7	70%	16	3	11	69%
D	NA	NA	NA	16	4	14	88%
Total	30	22	73%	64	7	49	77%

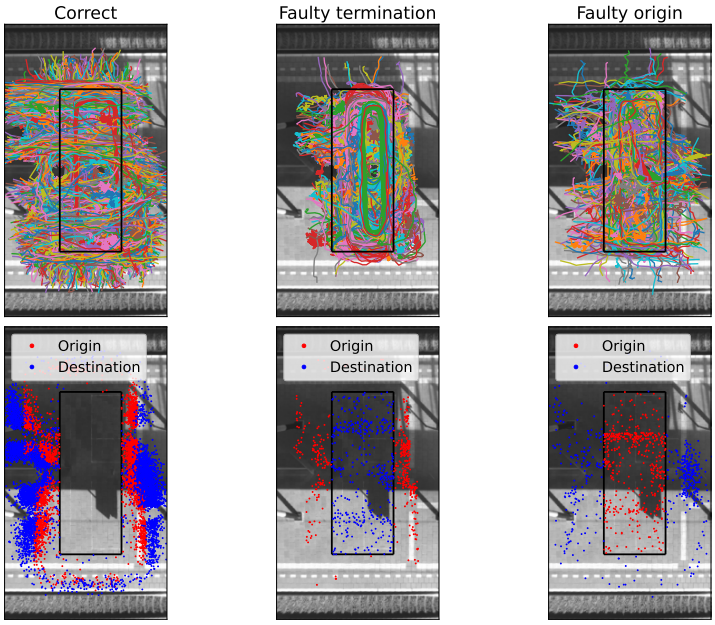


Figure 7 The recorded trajectories by the commercial setup during test 5. The upper row with figures reports the trajectories, partitioned per subset, and the bottom row reports the corresponding origin-destinations pairs. The black rectangle indicates the inner domain, S_{in}.

6. Discussion

In this contribution we presented a benchmarking suite for pedestrian tracking systems. The suite is light-weight and easily reproducible as it only contains a minimal set of 5 tests. The developed tests are tailored to take minimal efforts, taking typically less than two hours in total, while requiring only a dozen participants. Each test accurately targets the validation of one of the following key components of pedestrian tracking: line-based crowd flux estimation, global density estimation, individual position detection, and trajectory accuracy. The tests output quality factors expressed as single numbers. The combination of tests focuses on error-prone features like person-object recognition, and multi-sensor stitching. From a civil engineering standpoint, the tests reflect observables connecting with immediate awareness of a facility (1. instantaneous usage, 2. crowding distribution), as well as with longer-term efficiency and design (3. localization, 4-5. tracking, i.e. usage modes). Facility usage and crowd distributions can indicate potentially hazardous capacity issues and overcrowding in an early stage, while localization and tracking enable efficiency improvements such as separation of usage modes.

Together with the benchmarking suite we presented the benchmark results of two real-life pedestrian tracking systems, one commercial and one developed in academia. These test results, synthesized in Tab. 7, are meant as a reference of the state-of-the-art for new tracking installations and as a standard for novel tracking technologies. The higher accuracy of the commercial setup can be easily explained by its smaller measurement setup using fewer sensors. The high error in the density estimation test for the commercial

Table 7 Synthetic results of all tests for both tracking setups. We report for tests 1–5 the name, the metrics, and the scores of both setups. Additionally, we report underneath each score the conditions that impact the score the most.

Test	Name	Metric	TU/e setup	Commercial setup
1.	Line-based crowd flux estimation	$A^{(1)}$	95% at $J_A = 100$ ped/m and $d_{avg} \approx 0.8$ m	100% at $J_A = 30$ ped/m and $d_{avg} \approx 2$ m
2.	Global density estimation	$A^{(2)}$	97% at $\rho = 0.07$ ped/m^2	82% at $\rho = 0.85$ ped/m^2
3.	Individual position detection	$\sigma_{lin}^{(3)}$	7.1 cm with $N_{sens} = 12$	4.4 cm with $N_{sens} = 3$
		$\sigma_{loc}^{(3)}$	4.9 cm with $N_{sens} = 12$	4.4 cm with $N_{sens} = 3$
		$D^{(3)}$	98.8% with $N_{sens} = 12$	98.8% with $N_{sens} = 3$
		$L^{(3)}$	98.9% with $N_{sens} = 12$	99.4% with $N_{sens} = 3$
4.	Trajectory accuracy in controlled environment	$A^{(4)}$	73% at $S = 150$ m^2 and $d_{min} \approx 20$ cm	77% at $S = 50$ m^2 and $d_{min} \approx 20$ cm
5.	Trajectory accuracy in real-life environment	$A^{(5)}$	NA	79% at $S_{in} = 38$ m^2

setup is most likely caused by a systematic error due to faulty person-object differentiation. This emphasizes the great importance of background removal and proper sensor calibration. The benchmark results show us that optic-based tracking systems can estimate crowd fluxes and densities, and localize pedestrians with high accuracy. The biggest open challenge is Lagrangian time-tracking in case of complex and highly intertwined trajectories by pedestrians walking in close proximity. In this case the system were about 75% accurate.

Acknowledgements This work is part of the HTSM research program "HTCrowd: a high-tech platform for human crowd flows monitoring, modeling and nudging" with project number 17962, and the VENI-AES research program "Understanding and controlling the flow of human crowds" with project number 16771, both financed by the Dutch Research Council (NWO). The authors want to thank Dr. Antal Haans and Dr. Philip Ross for their effort in the establishment of the TU/e tracking setup.

References

[1] Eringa, P.: Prorail: Meer en snellere treinen. https://www.prorail.nl/nieuws/meer-en-snellere-treinen (2020). Accessed: 2021-07-28

[2] Daamen, W., Hoogendoorn, S.P.: Experimental research of pedestrian walking behavior. Transportation Research Record **1828**(1), 20–30 (2003). doi:10.3141/1828-03

[3] Seyfried, A., Steffen, B., Klingsch, W., Boltes, M.: The fundamental diagram of pedestrian movement revisited. Journal of Statistical Mechanics: Theory and Experiment **2005**(10), 10002 (2005). doi:10.1088/1742-5468/2005/10/P10002

[4] Kretz, T., Grünebohm, A., Kaufman, M., Mazur, F., Schreckenberg, M.: Experimental study of pedestrian counterflow in a corridor. Journal of Statistical Mechanics: Theory and Experiment **2006**(10), P10001 (2006). doi:10.1088/1742-5468/2006/10/P10001

[5] Moussad, M., Helbing, D., Garnier, S., Johansson, A., Combe, M., Theraulaz, G.: Experimental study of the behavioural mechanisms underlying self-organization in human crowds. Proceedings of the Royal Society B: Biological Sciences **276**, 2755–2762 (2009). doi:10.1098/RSPB.2009.0405

[6] Schadschneider, A., Klingsch, W., Klüpfel, H., Kretz, T., Rogsch, C., Seyfried, A.: Evacuation Dynamics: Empirical Results, Modeling and Applications. In: Extreme Environmental Events, pp. 517–550. Springer, New York, NY (2011). doi:10.1007/978-1-4419-7695-6_29

[7] Seitz, M.J., Köster, G.: Natural discretization of pedestrian movement in continuous space. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **86**(4) (2012). doi:10.1103/PhysRevE.86.046108

[8] Yamamoto, H., Yanagisawa, D., Feliciani, C., Nishinari, K.: Body-rotation behavior of pedestrians for collision avoidance in passing and cross flow. Transportation Research Part B: Methodological **122**, 486–510 (2019). doi:10.1016/j.trb.2019.03.008

[9] Corbetta, A., Bruno, L., Muntean, A., Toschi, F.: High statistics measurements of pedestrian dynamics. Transportation Research Procedia **2**, 96–104 (2014). doi:10.1016/J.TRPRO.2014.09.013

[10] Brščić, D., Zanlungo, F., Kanda, T.: Density and velocity patterns during one year of pedestrian tracking. In: Transportation Research Procedia, vol. 2, pp. 77–86 (2014). doi:10.1016/j.trpro.2014.09.011

[11] Zanlungo, F., Yücel, Z., Brščić, D., Kanda, T., Hagita, N.: Intrinsic group behaviour: Dependence of pedestrian dyad dynamics on principal social and personal features. PLoS ONE **12**(11) (2017). doi:10.1371/journal.pone.0187253

[12] Corbetta, A., Meeusen, J.A., Lee, C., Benzi, R., Toschi, F.: Physics-based modeling and data representation of pairwise interactions among pedestrians. Physical Review E **98**(6) (2018). doi:10.1103/PhysRevE.98.062310

[13] Willems, J., Corbetta, A., Menkovski, V., Toschi, F.: Pedestrian orientation dynamics from high-fidelity measurements. Scientific Reports 2020 10:1 **10**(1), 1–10 (2020). doi:10.1038/s41598-020-68287-6

[14] Pouw, C.A.S., Toschi, F., van Schadewijk, F., Corbetta, A.: Monitoring physical distancing for crowd management: Real-time trajectory and group analysis. PLOS ONE **15** (2020). doi:10.1371/journal.pone.0240963

[15] Boltes, M., Seyfried, A.: Collecting pedestrian trajectories. Neurocomputing **100**, 127–133 (2013). doi:10.1016/J.NEUCOM.2012.01.036

[16] Brščić, D., Kanda, T., Ikeda, T., Miyashita, T.: Person tracking in large public spaces using 3-D range sensors. IEEE Transactions on Human-Machine Systems **43**(6), 522–534 (2013). doi:10.1109/THMS.2013.2283945

[17] Seer, S., Brändle, N., Ratti, C.: Kinects and human kinetics: A new approach for studying pedestrian behavior. Transportation Research Part C: Emerging Technologies **48**, 212–228 (2014). doi:10.1016/j.trc.2014.08.012

[18] Yoshimura, Y., Sobolevsky, S., Ratti, C., Girardin, F., Carrascal, J.P., Blat, J., Sinatra, R.: An analysis of visitors' behavior in the louvre museum: A study using bluetooth data. Environment and Planning B: Planning and Design **41**(6), 1113–1131 (2014). doi:10.1068/b130047p

[19] Centorrino, P., Corbetta, A., Cristiani, E., Onofri, E.: Managing crowded museums: Visitors flow measurement, analysis, modeling, and optimization. Journal of Computational Science **53**, 101357 (2021). doi:10.1016/j.jocs.2021.101357

[20] Hong, H., De Silva, G.D., Chan, M.C.: Crowdprobe: Non-invasive crowd monitoring with wifi probe. In: Proceedings of the ACM Interactactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, p. 23. Association for Computing Machinery, New York, NY, USA (2018). doi:10.1145/3264925

[21] Georgievska, S., Rutten, P., Amoraal, J., Ranguelova, E., Bakhshi, R., de Vries, B.L., Lees, M., Klous, S.: Detecting high indoor crowd density with wi-fi localization: a statistical mechanics approach. Journal of Big Data **6** (2019). doi:10.1186/S40537-019-0194-3

[22] van den Heuvel, J., Thurau, J., Mendelin, M., Schakenbos, R., van Ofwegen, M., Hoogendoorn, S.P.: An application of new pedestrian tracking sensors for evaluating platform safety risks at swiss and dutch train stations. In: Hamdar, S.H. (ed.) Traffic and Granular Flow '17, pp. 277–286. Springer International Publishing (2019). doi:10.1007/978-3-030-11440-4_42

[23] Thurau, J., van den Heuvel, J., Keusen, N., van Ofwegen, M., Hoogendoorn, S.P.: Influence of Pedestrian Density on the Use of the Danger Zone at Platforms of Train Stations. In: Traffic and Granular Flow '17, pp. 287–296. Springer International Publishing (2019). doi:10.1007/978-3-030-11440-4_32

[24] Thurau, J., Keusen, N.: Influence of obstacles on the use of the danger zone on railway platforms. Collective Dynamics **5**(0), A84 (2020). doi:10.17815/cd.2020.84

[25] Microsoft Coorporation, Kinect for Xbox 360, Redmond WA, USA

[26] Kroneman, W., Corbetta, A., Toschi, F.: Accurate pedestrian localization in overhead depth images via height-augmented hog. Collective Dynamics **5** (2020). doi:10.17815/CD.2020.30

[27] Corbetta, A., Kroneman, W., Donners, M., Haans, A., Ross, P., Trouwborst, M., Van de Wijdeven, S., Hultermans, M., Sekulovski, D., Van der Heijden, F., Mentink, S., Toschi, F.: A large-scale real-life crowd steering experiment via arrow-like stimuli. Collective Dynamics **5**(0), 61–68 (2020). doi:10.17815/cd.2020.34

[28] Allan, D.B., Caswell, T., Keim, N.C., van der Wel, C.M., Verweij, R.W.: soft-matter/trackpy: Trackpy v0.5.0 (2021). doi:10.5281/zenodo.4682814

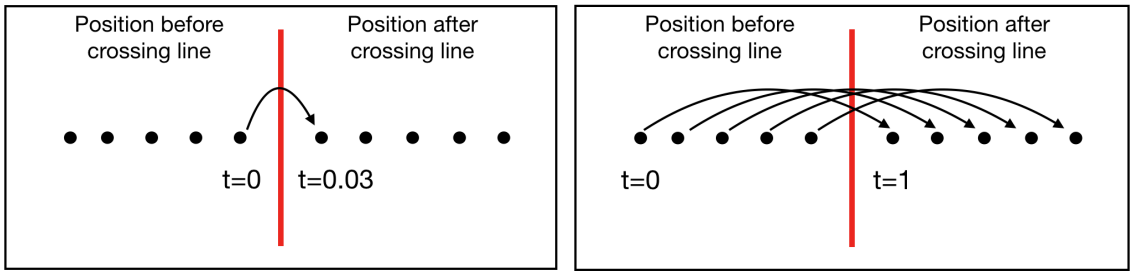
[29] Corbetta, A., Lee, C., Benzi, R., Muntean, A., Toschi, F.: Fluctuations around mean walking behaviors in diluted pedestrian flows. Physical Review E **95**(3) (2017)

[30] Liu, X., Song, W., Zhang, J.: Extraction and quantitative analysis of microscopic evacuation characteristics based on digital image processing. Physica A: Statistical Mechanics and its Applications **388**(13), 2717–2726 (2009). doi:10.1016/j.physa.2009.03.017

A. Counting algorithm

In this appendix we describe the algorithm that we used to quantify the number of pedestrians crossing a line in the TU/e setup for test 1.

A naive counting algorithm could verify whether two consecutive measurements land on opposite sides of a (virtual) reference line (see Fig. 8a). Such an approach is highly sensitive to measurement noise and could yield miscounts. In our test, we adopted a more robust approach determining a line crossing event based on multiple samples before and after the line. Specifically, we consider pairs of measurements along the same trajectory that are $\Delta T = 1$ s apart (i.e. 30 samples). A line crossing event was triggered when, within the ΔT time interval, the majority of pairs were located on different sides of the line (see Fig. 8b).



(a) (b)

Figure 8 Conceptual sketch of algorithms detecting the crossing of a virtual line (in red). Black dots indicate position measurements. (a) Conventional counting algorithm probing for two consecutive measurements on either side of the (virtual) crossing line. At $f = 30$ frames per seconds the measurements are $\Delta T = 0.03$ s apart. (b) Procedure to determine line crossings that we used in the TU/e setup for test 1. See explanation in App. A.