Transportation Research Part C 48 (2014) 212–228

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Kinects and human kinetics: A new approach for studying pedestrian behavior

Stefan Seer [a,b,*], Norbert Brändle [a], Carlo Ratti [b]

[a] Austrian Institute of Technology (AIT), Giefinggasse 2, 1210 Vienna, Austria
[b] MIT Senseable City Lab, Massachusetts Institute of Technology (MIT), 77 Massachusetts Avenue, 02139 Cambridge, MA, USA

ARTICLE INFO

Article history:
Received 11 September 2012
Received in revised form 30 May 2014
Accepted 14 August 2014
Available online 26 September 2014

Keywords:
Pedestrian simulation
Model calibration
Microsoft Kinect
People tracking
Ubiquitous sensing
Pervasive computing

ABSTRACT

Microscopic pedestrian simulation models can be used to investigate pedestrian movement at the urban block and building model scale. In order to develop, calibrate and validate such microscopic models, highly accurate and detailed data on pedestrian movement and interaction behavior (e.g. collision avoidance) is required. We present a data collection approach for studying pedestrian behavior which uses the increasingly popular low-cost sensor Microsoft Kinect. The Kinect captures both standard camera data and a three-dimensional depth map. Our human detection and tracking algorithm is based on agglomerative clustering of privacy-preserving Kinect depth data captured from an elevated view – in contrast to the lateral view used for gesture recognition in Kinect gaming applications. Our approach transforms local Kinect 3D data to a common world coordinate system in order to obtain human trajectories from multiple Kinects, which allows for a scalable and flexible capturing area. At a testbed with real-world pedestrian traffic we demonstrate that our approach can provide accurate trajectories from three Kinects with a Pedestrian Detection Rate of up to 94% and a Multiple Object Tracking Precision of 4 cm. Using a comprehensive dataset of 2674 captured human trajectories we calibrate three variations of the Social Force model. Data for model calibration and validation was recorded without any script and without actors behaving according to scripted situations. Various conditions have been covered in the dataset, such as walking at different densities, walking-stopping-walking, abrupt changes of direction and random movement. The results of our model validations indicate their particular ability to reproduce the observed pedestrian behavior in microscopic simulations.

1. Introduction

With 60% of the world's population projected to live in urban areas by 2030, crowd management and modeling is becoming an urgent issue of global concern (United Nations, 2012). A better understanding of pedestrian movement can lead to an improved use of public spaces, to the appropriate dimensioning of urban infrastructure such as airports, stations and commercial centers.

At the urban block and building model scale, predictions on crowd movement are usually being investigated using microscopic pedestrian simulation models. As specified in Hoogendoorn and Bovy (2004), modeling of pedestrian motion must

* Corresponding author at: Austrian Institute of Technology (AIT), Giefinggasse 2, 1210 Vienna, Austria.
E-mail addresses: stefan.seer@ait.ac.at, seer@mit.edu (S. Seer), norbert.braendle@ait.ac.at (N. Brändle), ratti@mit.edu (C. Ratti).

http://dx.doi.org/10.1016/j.trc.2014.08.012
0968-090X/

take into account human behavior on three levels. The *strategic level* describes the choice of general activity and trip purposes. The *tactical level* determines the route-choice and intermediate goals. The *operational level* computes the actual movement at each time instant towards the next goal and includes collision avoidance based on interactions with other pedestrians and the environment. This paper focuses on human behavior at the operational level.

Several microscopic modeling approaches exist to describe individual human behavior and collective phenomena on the operational level (see Duives et al. (2013) for an overview). The most common microscopic approaches noted in the scientific literature are cellular automata, discrete choice models and social force models. The cellular automata approach as described in Burstedde et al. (2001) composes a discretization of space using a square grid structure, where the cell sizes represent a single pedestrian. Also alternative grid structures such as hexagonal cells have been developed (Klein et al., 2010). While cellular automata models are generally considered to be effective in terms of computational performance, their practical application is somehow limited: one disadvantage is the limited accuracy of spatial representations (e.g. door widths are restricted to a multiple of the defined cell size) and pedestrian dynamics (i.e. pedestrian bodies are treated as incompressible). Spatially continuous approaches are computationally more expensive, but allow for modeling in greater detail. The social force approach (Helbing and Molnár, 1995) defines attraction and repulsion forces with respect to other humans and the environment, and has become one of the main approaches for modeling and simulating pedestrian movement on a microscopic level both in the scientific community and in commercial applications (e.g. VISWALK from PTV Group (2013) or SimWalk from Savannah Simulations (2013)). Discrete choice models (Antonini et al., 2006; Guo et al., 2010) use a representation where a pedestrian probabilistically selects the direction and speed of movement from a predetermined set of choices considering the dynamics of other pedestrians around. Some operational models extend the typical pedestrian movement in a way such that pedestrians include decisive behavior rather than just being reactive particles: Moussaïd et al. (2011) demonstrate a cognitive science approach based on behavioral heuristics by applying two simple cognitive procedures to adapt walking speed and direction. The model developed by Pellegrini et al. (2009) represents pedestrian interaction through minimizing an energy function for each individual.

An essential step in model development is the calibration with relevant datasets and the validation on different realistic scenarios. Model parameter estimation can be very complex due to large numbers of parameters or the limited availability of real-world datasets (see Rudloff et al., 2011). Some papers show that the calibrated models reproduce self organizing behavior such as lane formation (e.g. Helbing and Molnár, 1997; Moussaïd et al., 2010) and provide quantitative calibration results (Johansson et al., 2007). However, research in microscopic pedestrian simulation mostly concentrates on model development, while calibration and validation processes are often neglected. One way to calibrate microscopic pedestrian simulation models is to use aggregated data such as density-flow relationships (according to so-called fundamental diagrams) as described in Davidich and Köster (2012). Developing and calibrating models for detailed human interaction behavior (e.g. collision avoidance) requires *highly accurate* data on pedestrian movements, i.e. spatio-temporal motion trajectories as well as the analysis of *all* people in a given scene.

In the pedestrian simulation community, such data is traditionally collected by manually annotating the positions of people in individual frames of recorded video data of highly frequented areas (Antonini et al., 2006; Berrou et al., 2007). Sometimes additional attributes such as age or gender are assigned to people during the annotation process. Manual annotation, however, is tedious and time-consuming, in particular for dense scenes with many pedestrians. This limits the amount of data that can be analyzed. Semi-automated video annotation approaches can facilitate the generation of motion data by providing automatically computed position predictions which support manual annotation (Plaue et al., 2011; Johansson and Helbing, 2010). Semi-automated video annotation can provide significant speedup compared to purely manual annotation, but is still time-consuming. In some experimental setups for pedestrian simulation modeling, participants of the experiments are equipped with distinctive wear such as colored hats for better identification. Naturally, such distinctive wear makes automated extraction of trajectories a relatively easy task. The free software *PeTrack* presented in Boltes et al. (2010) has been applied on video recordings of a bottleneck experiment. The automatic tracking approaches of Hoogendoorn and Daamen (2003) and Hoogendoorn and Daamen (2005) collected trajectory data in a narrow bottleneck and a four-directional crossing flow experiment. Controlled experiments allow the setting of environmental conditions that are hard to observe in real world circumstances as in Daamen and Hoogendoorn (2012), where emergency settings were reenacted including acoustic and visual signals. However, such setups only allow for a limited sample size and include a significant bias in the data since participants are usually aware of being observed.

Large scale and real-world data on human motion can be obtained from video only when applying tools for automatic vision-based people detection and tracking. Vision based tracking has seen considerable progress in recent years, with current systems able to track people through long and challenging sequences (Pellegrini et al., 2009). There exist many approaches for multi-camera people tracking assuming overlapping fields of view, e.g. Fleuret et al. (2008), Eshel and Moses (2010), Sternig et al. (2011). Approaches dealing with camera networks with non-overlapping fields of view include Javed et al. (2007), Loy et al. (2009) and Pflugfelder and Bischof (2010). The advanced computer vision methods are mostly not concerned with highly accurate tracking (in the sense of giving a position estimate within centimeters of the actual position), but are more concerned with working under difficult conditions (occlusions, clutter, lighting, ...), which limits their reliability and makes them less appropriate for measurement tasks. The computer vision community provides publicly available benchmark sequence datasets with manually or semi-automatically annotated data for testing common surveillance tasks, such as pedestrian detection, tracking and behavior analysis, e.g. PETS (PETS, 2009) and VIRAT (Oh et al., 2011). Many of these sequences were captured with people *acting* behavior according to a script, often also due to legal and privacy issues.

The 3DPeS dataset of Baltieri et al. (2011) provides a benchmark dataset primarily for people re-identification, where subjects were notified of the presence of cameras, but not coached or instructed in any way. Benchmark datasets where subjects were not notified of the presence of cameras are provided in the BIWI walking dataset of ETH Zurich (Pellegrini et al., 2009) and the 'crowds-by-example' dataset of University of Cyprus Computer Graphics Lab (2011). These datasets were generated by manual or semi-automated annotation for the sake of ground truth, and were used, for example, to demonstrate novel people tracking algorithms improved with Social Force Models (Pellegrini et al., 2009) or algorithms for automatically inferring groups with Social Force Models (Sochman and Hogg, 2011). Notably, these algorithms are published by renowned computer vision labs and critically rely on accurate trajectory data of every person in the scene, and are still *demonstrated only on the manually or semi-automatically annotated ground truth trajectory data.*

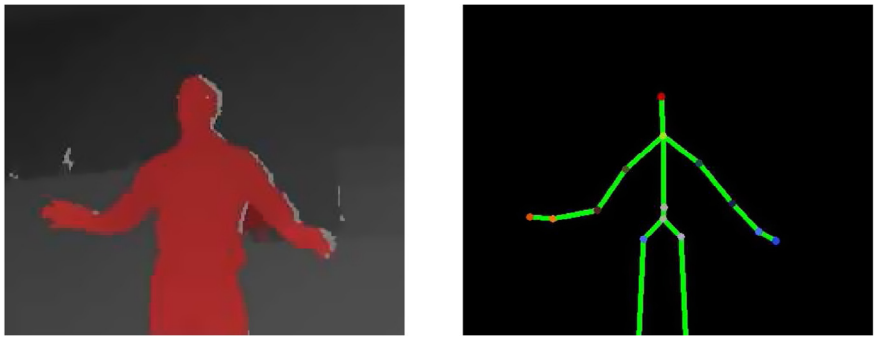
Mutual occlusions of people are one of the major challenges for automatic multiple people tracking from oblique camera views, and they can be minimized or avoided when capturing the scenes from a top-down view. The Edinburgh Informatics Forum Pedestrian Database (Bob Fisher, 2010) provides a comprehensive set of trajectories computed in video from a fixed overhead camera 23 meters from the floor. This elevated height covers a large area, and the authors in (Bob Fisher, 2010) note that there are usually only a few individuals in each video frame (1 person in 46 % of the frames, four or less persons in 93 % of the frames), and that scenes with many people (events) cause difficulties in the tracker. The dataset has been used to demonstrate surveillance algorithms, in particular to detect unusual behavior, for example in Calderara et al. (2011).

In this paper we propose an approach for calibrating microscopic pedestrian simulation models with highly accurate and comprehensive trajectory data on individual pedestrian movement using Microsoft Kinects mounted in top-down positions. The Kinect is an inexpensive motion sensing input device which was originally developed for the Xbox 360 video game console and delivers RGB camera images and a 3D depth map (Microsoft Corp., 2012a). It was originally designed to accurately detect three dimensional positions of body joints (Shotton et al., 2011) and to estimate human pose (Girshick et al., 2011). Fig. 1 illustrates the *skeletal tracking* which is the key component of the video game user interface. With its built-in functionality, the Kinect can detect up to six people (two of them using the skeletal tracking) provided that all persons face the sensor in frontal view with their upper bodies visible. Since its market introduction in 2010, the Kinect has also been used in a broad variety of other research fields: Noonan et al. (2011) showed the use of the Kinect for tracking body motions in clinical scanning procedures. Animation of the hand avatar in a virtual reality setting by combining the Kinect with wearable haptic devices was developed in Frati and Prattichizzo (2011). The Kinect was used in Izadi et al. (2011) to create detailed three dimensional reconstructions of an indoor scene. Weiss et al. (2011) presented a method for human shape reconstruction using three dimensional and RGB data provided by the Kinect. Choi et al. (2013) demonstrate their method for multiple people tracking from a moving camera using Kinect datasets.

To the best of our knowledge, the Microsoft Kinect has not yet been used to obtain data for modeling pedestrian motion behavior. The built-in skeletal tracking cannot be directly used for accurately measuring pedestrian movement.

In this paper we provide a contribution for pedestrian researchers by presenting an alternative approach using multiple Microsoft Kinects for obtaining highly accurate tracking data from an elevated view for pedestrian modeling. The approach processes only information from the 3D depth sensor, thus avoiding any legal and privacy issues arising when observing people in real-world scenarios with imaging in the visual spectrum. We demonstrate the high accuracy of the trajectories in a real world setup and show how such an automatically obtained set of more than 2600 trajectories is used to calibrate and compare microscopic pedestrian simulation models. The overall added value with respect to previous approaches is the feasibility to automatically obtain large and accurate sets of human movement data at low cost paving the way for many pedestrian researchers to better calibrate and validate their simulation models.

We combined three Kinect sensors and collected a large dataset on pedestrian movement inside the Massachusetts Institute of Technology (MIT)'s Infinite Corridor, the longest hallway that serves as the most direct indoor route between the east and west ends of the campus and is highly frequented by students and visitors. Fig. 2a shows the area identified for the data



(a) (b)

Fig. 1. Microsoft Kinect provides the depth data stream (a), with the detected person in red and different gray levels encoding the depth information, and skeletal tracking (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(a) (b)

Fig. 2. MIT's Infinite Corridor with (a) the observed area (green) and (b) the Kinect setting on the ceiling. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

collection in this work, and Fig. 2b shows the Kinect sensors mounted at the ceiling. In our setting, a single Kinect sensor is capable of covering an area of roughly 2 m×2 m. Since we have suspended three Kinect sensors from the ceiling, human movement behavior in an overall scanning area of around 12m^2 could be measured. In order to observe various pedestrian behaviors we performed different walking experiments in this environment.

This paper is structured as follows: Section 2 outlines the setting for measuring human motion data using the Kinect, describes the calibration process to derive world coordinate data from the Kinect sensors and the algorithms for detecting and tracking of humans using multiple Kinects. Section 3 provides evaluation results demonstrating the tracking performance in the described setting. Section 4 describes the walking experiments and data collection at MIT's Infinite Corridor. We describe how these data sets can be used for the calibration of pedestrian models and provide results from the calibration and validation of three simulation models. Section 5 concludes the results and gives an outlook for further research.

2. Human detection and tracking

Accurate movement data of all humans in the scene are of vital importance for the calibration and validation of microscopic pedestrian simulation models. The Kinect provides a sequence of standard RGB color frames and a 3-dimensional depth image for each frame. The depth image of a scene indicates the distance of each picture element of that particular scene from the Kinect. Depth images and RGB color images are both accessible with the *Kinect for Windows SDK* by Microsoft Corp. (2012b). Fig. 3 illustrates a snapshot of the depth image, the RGB image and a combination of depth and RGB from three Kinects mounted at a height of 4.5 meters and a top view position in the MIT's Infinite Corridor. With this setup a section of 6 meters of the corridor can be captured. Note that the glass case introduces a significant amount of artifacts due to specular reflections. In order to meet privacy concerns – most of the observed persons are not aware of any data collection experiment – our approach does not process RGB information from the visible spectrum.

In order to compute pedestrian trajectories from depth image sequences of multiple Kinects, we (1) map depth information from individual Kinect sequences into a common world coordinate system, (2) group depth information from a single Kinect in the world coordinate system into individual pedestrians and (3) track the pedestrians to obtain trajectories throughout the sensing areas of multiple Kinect sensors. These three steps are described in the following subsections.

2.1. Obtaining world coordinates

A Kinect sensor S_k from a set of K devices generates a time series of 640×480 depth pixel images. Each depth image encodes a set of valid three-dimensional points $\mathbf{x}_{c_i} = [x_{c_i}\ y_{c_i}\ z_{c_i}]^T$, with $i \leqslant 640 \times 480$, in the local Kinect 3D camera coordinate system, computed with the value of the focal length f provided by Microsoft Corp. (2012b). The physical constraints of the Kinect 3D-measurement setup limit the range of z_{c_i} within which reliable depth data can be computed to a maximum distance of 4 meters. Objects which are located more than 4 meters away from the sensor can not be captured.

A human trajectory $\mathcal{T}$ is denoted as a sequence of N four-dimensional vectors

$$\mathcal{T} = \{[t_i\ x_{w_i}\ y_{w_i}\ z_{w_i}]^T\}_{i=1\ldots N,} \tag{1}$$

where the vectors are composed of a timestamp t_i and a 3D position $\mathbf{x}_{w_i} = [x_{w_i}\ y_{w_i}\ z_{w_i}]^T$ in a common world coordinate system: For a trajectory to represent people walking throughout the sensing areas of multiple Kinect sensors, the points of the local 3D coordinate systems of the mounted Kinect sensors must first be mapped to the world coordinate system.

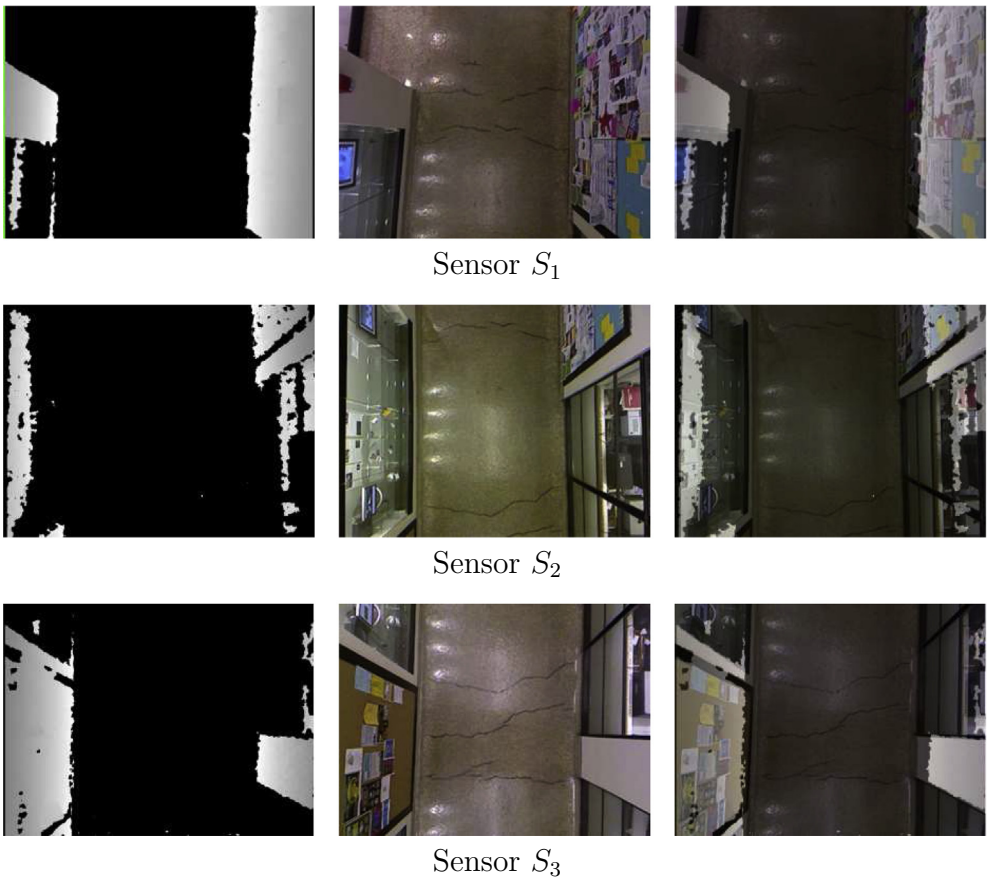


Fig. 3. Kinect sensor field of view; raw depth data stream (left), RGB stream (middle) and both data streams in an overlay (right).

The actual point mapping between the coordinate system of sensor S_k and the world coordinate system is represented by a rigid transformation, composed of a translation vector $\mathbf{t}_k$ between the two origins of the coordinate systems and a 3×3 rotation matrix $\mathbf{R}_k$ such that

$$\mathbf{x}_{w_i} = \mathbf{R}_k \mathbf{x}_{c_i} + \mathbf{t}_k. \tag{2}$$

Note that we do not model nonlinear lens distortion. As elaborated in Konolige and Mihelich (2012), the Kinect lenses are already very good compared to, for example, typical webcams, with a reprojection error (deviation of the camera from the ideal pinhole model) of 0.34 pixel for the IR camera. Konolige and Mihelich (2012) show that modeling lens correction could reduce the reprojection error by factor of 3. Lens correction is useful for RGB images with subpixel accuracy applications. Kinect IR depth images, however, are already smoothed over a neighborhood of pixels by the correlation window, hence correcting a third of a pixel would not significantly change the depth result.

The three parameter values for translation $\mathbf{t}_k$ of sensor S_k and its three rotation angles in $\mathbf{R}_k$ are determined by a set of M point matches $< \mathbf{x}_{w_i}, \mathbf{x}_{c_i} >, i \in M$ and subsequently minimizing the error

$$E = \sum_{i=1}^{M} \left\| \mathbf{x}_{w_i} - \mathbf{R}_k \mathbf{x}_{c_i} - \mathbf{t}_k \right\|^2 \tag{3}$$

by solving an overdetermined equation system as described in Forsyth and Ponce (2002).

We determine the M point matches $< \mathbf{x}_{w_i}, \mathbf{x}_{c_i} >$ in world coordinates $\mathbf{x}_{w_i}$ manually from the depth images. Since only depth information and no visual information is available for the sensed area, sensor calibration must be based on pre-determined calibration objects with well-defined depth discontinuities. Our sensor calibration setup is composed of a rectangular piece of cardboard placed on a tripod with a height of 1587 mm from the ground. We marked multiple reference points on the ground assuring that they are equally distributed within the field of view. One by one, the tripod with the mounted cardboard on top was placed perpendicular to a single reference point and the depth data was recorded for a couple of seconds. The reference points in world coordinates $\mathbf{x}_{w_i}$ are determined as the center of gravity of the extracted cardboard corners in the depth images. The raw depth data including the reference points and the results of the calibration for all sensors are

shown in Fig. 4. Table 1 shows that the Root-Mean-Square Error (RMSE) between the reference points in the world coordinates and the reference points in camera coordinates transformed with (2) lies within the range of a few centimeters.

2.2. Detection algorithm

Let $\mathcal{D}$ denote the set of points $\mathbf{x}_{w_i}$ obtained by applying the rigid transform (2) to the 3D camera coordinates from the Kinect depth images. The objective of human detection is to extract from $\mathcal{D}$ connected sets of points belonging to a person and to represent the person with a point $\mathbf{x}_{p_i}$. Human tracking associates detections of individuals over time. Human detection is composed of the following steps:

1. **Data reduction by background subtraction**. Identifying a set of points which do not change or only change slowly over time – the background – supports the segmentation of walking persons from other objects and reduces the number of depth points to be processed. This can be achieved by classic background subtraction techniques from the domain of video analysis, e.g. the adaptive background modeling with Gaussian Mixture Models described in Stauffer and Grimson (2000). In our particular case of the Infinite Corridor, the background model is handcrafted, since the locations of background objects such as walls are well-known in advance.
2. **Data reduction by cutoff.** The cutoff step first removes all 3D points which remain after background subtraction with height z_{w_i} larger than a tall person's height, e.g. 2.1 meters for adults, and all 3D points with height z_{w_i} smaller than a typical upper body region, e.g. 1.5 meters. The second cutoff value determines the minimal height of detectable persons, and is necessary to exclude noisy measurements of objects near the floor. Applying the cutoff values to z_{w_i} results in a subset $\mathcal{D}'$.

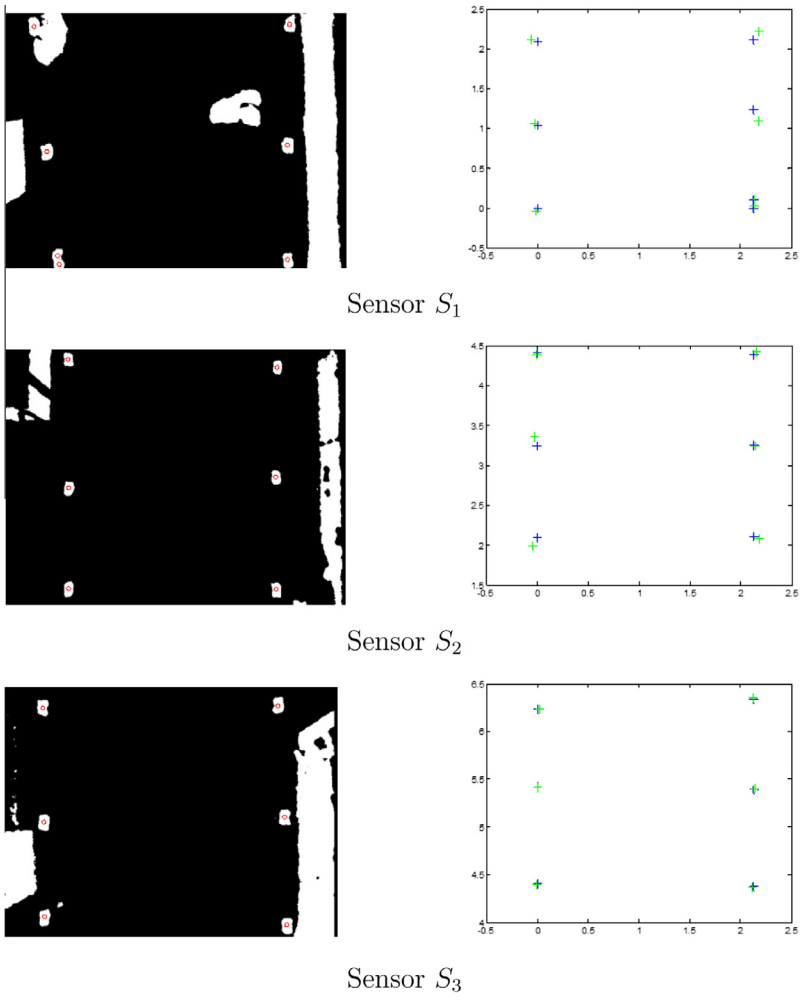


Fig. 4. Left column - Raw data from sensor including reference points (red circles); Right column - sensor calibration results with measured reference (blue) and estimated (green) points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Accuracy of calibration computed on reference points.

	Sensor S_1	Sensor S_2	Sensor S_3
RMSE	64 mm	67 mm	19 mm

3. **Hierarchical clustering on the reduced set.** In order to group the points $\mathcal{D}'$ into natural clusters corresponding to individual persons, we first build a cluster tree by agglomerative clustering with the complete-linkage algorithm (Duda et al., 2001). For computational reasons we randomly select a subset $\mathcal{D}''$ of R points from $\mathcal{D}'$ for clustering, where typically $R = 500$. The complete-linkage algorithm uses the following distance $d(\mathcal{D}_i'', \mathcal{D}_j'')$ to measure the dissimilarity between subsets of $\mathcal{D}''$:

$$d(\mathcal{D}_i'', \mathcal{D}_j'') = \max_{\substack{\mathbf{x} \in \mathcal{D}_i'' \\ \mathbf{x}' \in \mathcal{D}_j''}} \|\mathbf{x} - \mathbf{x}'\|, \tag{4}$$

with $\|\cdot\|$ as the Euclidean distance. Using metric (4) avoids elongated clusters and is advantageous when the true clusters are compact and roughly equal in size (Duda et al., 2001). All leaves at or below a node with a height less than a threshold are grouped into a cluster, where the threshold is based on a typical human shoulder width, e.g. 0.6 meters.
4. **Grouping of $\mathcal{D}'$ and cleanup.** All available observation points of $\mathcal{D}'$ are assigned to a cluster, given that they are sufficiently close to the cluster center. Otherwise they are removed. Small clusters which originate from noise or people on the border of the field of view are removed.
5. **Identifying a cluster representative.** For every cluster $\mathcal{D}_i''$, the point $\mathbf{x}_{\mathrm{p}_i}$ representing the pedestrian location of a trajectory (1) is selected as the point with the 95th percentile of the height z_{w_i} in $\mathcal{D}_i''$, defined as the person's height.

This process provides robust people detections of all individuals in a single depth image.

2.3. Tracking over multiple sensor views

In order to establish correspondences between consecutive detections and obtain trajectories $\mathcal{T}$ as defined in (1), we perform global tracking of the people detections $\mathbf{x}_{\mathrm{w}_i}$ in the world coordinate system by a simple nearest neighbor matching based on position predictions linearly extrapolating individuals' positions of the previous n detections. We search for the nearest detection in the neighborhood within a spatial and temporal threshold.

While other applications use more complex approaches for object tracking (see Berclaz et al. (2011) for an overview), we take advantage of the high rate of 30 frames per second provided by the Kinect. We experimentally determined $n = 5$, taking into account a sufficient number of observations for obtaining a reliable estimate of the predictions. Fig. 5 illustrates the tracking results of a short sequence as red lines, superimposed to raw depth data of a Kinect sensor.

The algorithm effectively tracks pedestrian heads, and the question might arise whether the head of a pedestrian is the most stable object to track. Increased variability of vertical head translation and pitch rotation of the head with respect to pitch trunk rotation was observed only for fast speeds (> 1.4 m/s) during experimental setups in Hirasaki et al. (1999). Additional variation might occur due to yaw rotation of the head for persons looking around while walking or standing in the observed area. Hence, shoulders as viewed from above might be a bit more stable when compared to head variations, in particular for fast speeds. Despite this, we prefer detection of head center points in the 3D data from the Kinect as the more straightforward and less error-prone approach, supported by the evaluation results of Section 3.

Having the Kinect sensors in a slightly overlapping setting avoids unobserved regions without any information on the location of pedestrians, and consequently enables more robust tracking. Fig. 6 illustrates the results of our automatic tracking approach in a scene with three persons (current position denoted by dashed circles) based on detections from three slightly overlapping views. When using overlapping sensors a single person potentially creates more than one detection in world coordinate space. This leads to artifacts within the resulting trajectories in form of small fluctuations (see magnification in Fig. 6). We therefore resample each individual trajectory $\mathcal{T}$ as defined in (1) comprising N detections to a trajectory $\mathcal{T}'$ with M vectors defined as

$$\mathcal{T}' = \{[t_i' \; x'_{\mathrm{w}_i} \; y'_{\mathrm{w}_i}]^T\}_{i=1\ldots M}, \tag{5}$$

where $t_i\prime = t_1, t_1 + \Delta t', t_1 + 2\Delta t', \ldots t_N$ with a sample rate $\Delta t' = 0.1$ seconds. Note that since the approaches for pedestrian modeling and calibration used in this work (see Section 4) make use of trajectories in 2-dimensional space only, we can omit z'_{w_i} in (5). The corresponding x'_{w_i} and y'_{w_i} at the regular time intervals $t_i\prime$ are obtained by cubic spline approximation according to de Boor (2001) using a smoothing parameter $p = 0.98$, which smoothes out the mentioned fluctuations and other small local variations along the trajectory (see black curve within the magnified box in Fig. 6).

Fig. 5. Kinect depth raw data in 3D (walls are light gray and detected objects are dark gray) with automatically obtained trajectories (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

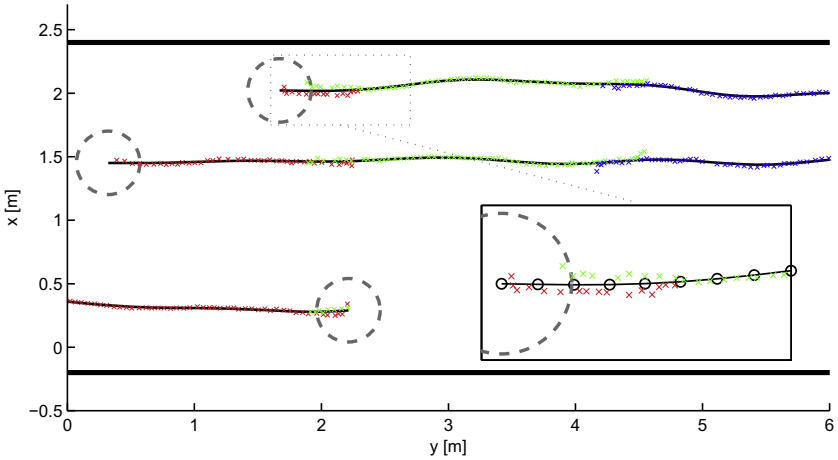


Fig. 6. Automatic tracking of three persons (current position denoted by dashed circles) based on detections from three slightly overlapping views (crossings in red, green and blue) with the resulting smoothed trajectories (black curve with circles, within the magnified box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Tracking evaluation

Real data for pedestrian simulation calibration is often confined to trajectories which have been manually extracted from video data sets. The reason is that the required accuracy of the trajectories is very high, and often only manually extracted trajectories can fulfill such accuracy requirements. It is thus necessary to compare the output of the Kinect pedestrian track-

ing described above with the "gold standard" of manually generated trajectories in order to have an idea how suitable automatic collection of really large data sets are.

A human observer annotated the locations of all individuals in single frames using the raw depth sensor data from a single Kinect sensor. While the Kinect's depth data does not allow for identifying persons, the body shape of individuals is still recognizable. Our evaluation data is composed of two trajectory sets: the first data set comprises 15578 frames ($\approx$ 590 s) with pedestrian flows of low to medium density, i.e. up to 0.5 persons/m^2, and a total number of 128 persons. The second sequence includes 251 frames ($\approx$ 12 s) with a total number of 21 persons and comparably higher densities of up to 1 person/m^2. Note that the stated densities are averaged over the covered area. Thus, the data do contain local densities well above 1 person/m^2, for instance in case that several pedestrians stand or walk closer together although a larger area is available. Fig. 7 illustrates a single frame from the second dataset.

In the first step of the evaluation, every automatically computed trajectory $\mathcal{T}$ is assigned to a ground truth trajectory $\mathcal{T}^G$ by minimizing a trajectory distance metric. Quantifying the pairwise trajectory dissimilarity in a distance metric is not trivial due to the usually different number of points. Here we used the discrete Fréchet distance (Eiter and Mannila, 1994). Following an informal interpretation, the Fréchet distance between two trajectories is the minimum length of a leash that allows a dog and its owner to walk along their respective trajectories, from one end to the other, without backtracking. Taking into account the location and ordering of points along the trajectories, the Fréchet distance is well-suited for the comparison of trajectories and is less sensitive to outlier points than alternatives for arbitrary point sets such as the Hausdorff distance.

As a result of the trajectory assignment we derive a set of P matching trajectory pairs for a time stamp t. Any remaining automatically computed trajectories which could not be matched are considered as false positives. Similarly, any remaining ground truth trajectories which could not be matched are considered as misses. Fig. 8a shows the results based on trajectories from the second sequence. Our dataset produced zero false positives and one miss. It was seen in the data that this missed person was smaller than the defined cutoff value of 1.5 meters. In order to quantify the position error for all correctly tracked objects over all frames, we use the Multiple Object Tracking Precision (MOTP) as described in Bernardin and Stiefelhagen (2008), which is defined as

$$Q_{\mathrm{MOTP}} = \frac{\sum_{i,t} \left\| \mathbf{x}_{\mathbf{i},\mathbf{t}} - \mathbf{x}^{\mathbf{G}}_{\mathbf{i},\mathbf{t}} \right\|}{\sum_t c_t}, \tag{6}$$

where c_t is the number of matches found at time t. For each match i, the Euclidean distance between the automatically computed trajectory point $\mathbf{x}_{\mathbf{i},\mathbf{t}}$ and the ground truth trajectory point $\mathbf{x}^{\mathbf{G}}_{\mathbf{i},\mathbf{t}}$ at time t is computed.

The evaluation results for our detection and tracking approach are shown in Table 2 and reveal that the localization errors in terms of MOTP are only within a few centimeters for both sequences. Fig. 8b illustrates the evolvement of MOTP over short time intervals based on the second sequence: for each time interval $\Delta t = 0.5$ seconds, the tracking precision $Q_{\mathrm{MOTP}^{\Delta t}}$ is computed equivalently to (6). This analysis reveals that tracking accuracy is stable and independent of the number of people. Fig. 9 illustrates the distribution of the Euclidean distances $\left\| \mathbf{x}_{\mathbf{i},\mathbf{t}} - \mathbf{x}^{\mathbf{G}}_{\mathbf{i},\mathbf{t}} \right\|$ according to (6) for all matches i between corresponding automatic and ground truth trajectory points for both sequences.

The Pedestrian Detection Rate (PDR) measures the rate at which tracked pedestrians are matched to the ground truth. The value of PDR varies between 0 and 1. While 0 means poor pedestrian detection, 1 means that all ground truth pedestrians are matched. The metric is given by

Fig. 7. Kinect depth raw data in 3D (gray) with manually annotated head positions of individuals (red circles). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

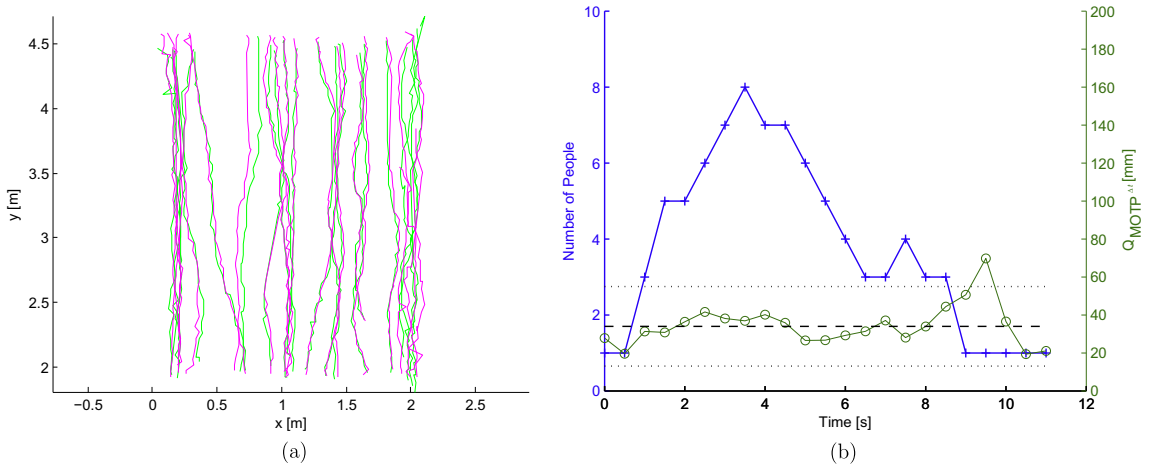


Fig. 8. Tracking performance evaluation including 21 persons with up to 1 person/m² based on (a) ground truth (green) and automatic trajectories (magenta) and (b) the temporal aspects of the Multi Object Tracking Precision (MOTP) with average (dashed line) and standard deviation (dotted line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Tracking evaluation results, showing Pedestrian Detection Rate (PDR) and Multi Object Tracking Precision (MOTP), and its standard deviation in brackets.

	Q_{PDR}[%]	Q_{MOTP}[mm]
Sequence 1	96.20 (12.39)	41.3 (30.2)
Sequence 2	93.86 (17.64)	34.0 (21.0)

$$Q_{PDR} = \frac{TP}{TP + FN}, \tag{7}$$

where the number of matched ground truth pedestrians is denoted by true positives TP. False negatives FN state the number of missing detections. Table 2 provides the evaluation results for our detection and tracking approach. Based on the PDR, our approach performs well on both sequences, with detection rates above 94%.

For Sequence 1 the PDR's for 115 out of 127 tracked persons are higher than 90% whereas the remaining 12 persons are lying in the range between 26.5% to 88.6%. For Sequence 2 the PDR's for 19 out of 21 persons are above 90% with the remaining 2 persons being in the range between 36.2% and 46.3%.

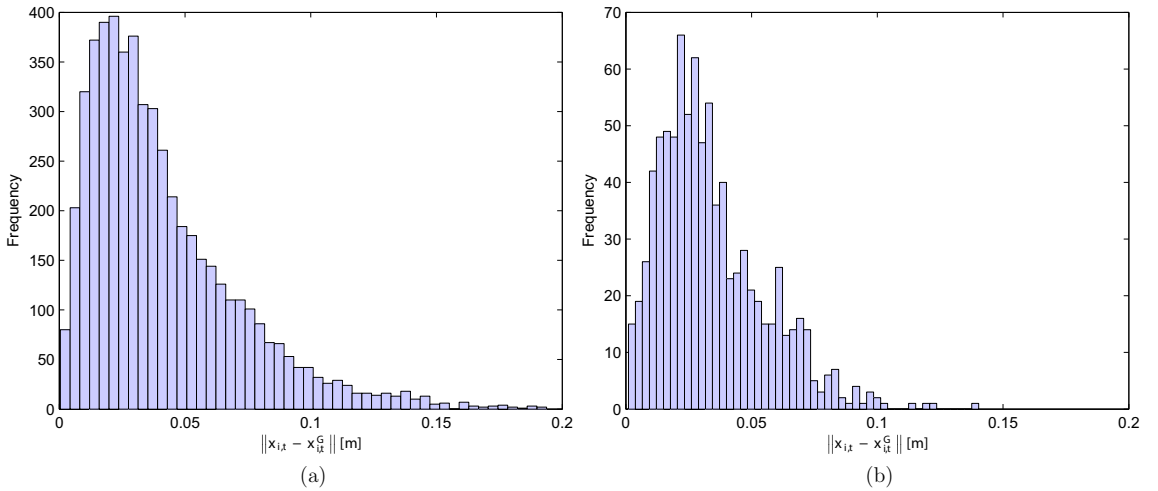


Fig. 9. Tracking evaluation results, showing the distribution of the Euclidean distance between the corresponding automatic trajectory point $\mathbf{x_{i,t}}$ and the ground truth trajectory point $\mathbf{x^G_{i,t}}$ for (a) Sequence 1 and (b) Sequence 2.

4. Calibration of pedestrian models

The people trajectories automatically obtained by the Kinect people detection and tracking described above have a positioning error of only a few centimeters with respect to manual annotation (see Table 2). Thus, a set of mounted and calibrated Kinects can quickly and easily produce large datasets of accurate empirical observations. Such movement datasets covering a variety of walking behavior are exactly what microscopic pedestrian models must rely on in order to be realistic.

We performed a variety of walking experiments at the MIT's Infinite Corridor described in Section 2 while capturing depth image sequences of the three Kinect sensors. Applying the people tracking algorithm on the collected Kinect raw depth data produced a comprehensive amount of robust trajectories providing the necessary information for calibrating different types of microscopic pedestrian simulation models. We present experimental results of comparing three variations of the social force model (see Helbing and Molnár, 1995) based on our data collected for calibrating these models.

4.1. Walking experiments

The walking experiments were performed under real world conditions, meaning that the individuals traversing MIT's Infinite Corridor had no information about being observed. The main task was to calibrate different microscopic pedestrian simulation models on the operational level with relatively simple scenarios, which allow to neglect the tactical level such as route choice.

In the first walking experiment, a person standing in the center of the observed area served as an obstacle for passing people. The 685 trajectories of this setting were recorded during a period of approximately 28 min (see Fig. 10a). The second walking experiment includes "normal" walking behavior without any external influence for a time span of around one hour. The 1989 trajectories computed with our Kinect approach are illustrated in Fig. 10b. The red and blue trajectories in Fig. 10a and b represent the two walking lanes in opposite directions which people form most of the time. From thorough investigation of both datasets we can confirm that various conditions are covered, including (but not exclusively) walking at different densities, walking-stopping-walking, abrupt changes of direction and random movement.

Fig. 11a and b show the walking speed histograms computed from the trajectories of the two calibration data sets (the velocity of the person acting as an obstacle in experiment 1 is filtered out). Fitted parameters of a Gaussian function to the data set result in a mean speed of 1.33 m/s and a standard deviation of 0.26 m/s. Experiment 2 shows similar results for the walking speed distribution with a mean speed of 1.30 m/s and a standard deviation of 0.33 m/s. Our results are well in line with findings in the scientific literature. For instance, the literature review in Daamen and Hoogendoorn (2006)

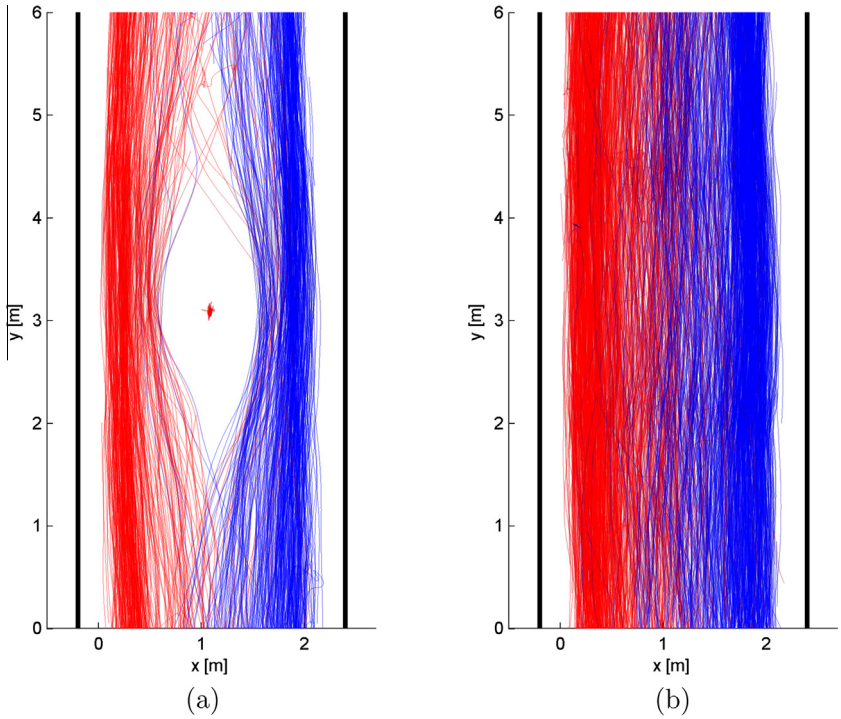


Fig. 10. Trajectories for the calibration of pedestrian models automatically retrieved from (a) experiment 1 and (b) experiment 2 (walking directions are encoded in red and blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

revealed that the walking speed of an individual appears to follow a normal distribution with an estimated mean of 1.34 m/s and a standard deviation of 0.37 m/s.

4.2. Investigated Social Force model variants

Given that the movement of a person depends on velocity and hence on acceleration, the principle of the Social Force model aims at representing individual walking behavior as a sum of different accelerations as

$$\mathbf{f}_\alpha(t) = \frac{v_\alpha^0 \mathbf{e}_\alpha - \mathbf{v}_\alpha}{\tau_\alpha} + \sum_{\beta \neq \alpha} \mathbf{f}_{\alpha\beta}(t) + \sum_i \mathbf{f}_{\alpha i}(t). \tag{8}$$

The acceleration $\mathbf{f}_\alpha$ at time t of an individual α towards a certain goal is defined by the desired direction of movement $\mathbf{e}_\alpha$ with a desired speed v_α^0. Here, the current velocity $\mathbf{v}_\alpha$ is adapted to the desired speed v_α^0 within a certain relaxation time τ_α. The movement of a pedestrian α is influenced by other pedestrians β which is modeled as a repulsive acceleration $\mathbf{f}_{\alpha\beta}$. A similar repulsive behavior for static obstacles i (e.g. walls) is represented by the acceleration $\mathbf{f}_{\alpha i}$. For notational simplicity, we omit the dependence on time t for the rest of the paper.

From the set of different formulations of the Social Force model available in the scientific literature, we compare three variations of the Social Force model based on the general formulation (8).

Model A: The first model from Helbing and Molnár (1995) is based on a circular specification of the repulsive force given as

$$\mathbf{f}_{\alpha\beta}^A = a_\alpha e^{-\frac{\left(r_\alpha + r_\beta - \left\|\mathbf{d}_{\alpha\beta}\right\|\right)}{b_\alpha}} \frac{\mathbf{d}_{\alpha\beta}}{\left\|\mathbf{d}_{\alpha\beta}\right\|}, \tag{9}$$

where r_α and r_β denote the radii of pedestrians α and β, and $\mathbf{d}_{\alpha\beta}$ is the distance vector pointing from pedestrian α to β. The interaction of pedestrian α is parameterized by the strength a_α and the range b_α, where their values are determined in the model calibration process.

Model B: The second model uses the elliptical specification of the repulsive force as described in Helbing and Johansson (2009) determined by

$$\mathbf{f}_{\alpha\beta}^B = a_\alpha e^{-\frac{w_{\alpha\beta}}{b_\alpha}} \frac{\mathbf{d}_{\alpha\beta}}{\left\|\mathbf{d}_{\alpha\beta}\right\|}, \tag{10}$$

where the semi-minor axis $w_{\alpha\beta}$ of the elliptic formulation is given by

$$w_{\alpha\beta} = \frac{1}{2}\sqrt{\left(\left\|\mathbf{d}_{\alpha\beta}\right\| + \left\|\mathbf{d}_{\alpha\beta} - \left(\mathbf{v}_\beta - \mathbf{v}_\alpha\right)\Delta t\right\|\right)^2 - \left\|\left(\mathbf{v}_\beta - \mathbf{v}_\alpha\right)\Delta t\right\|^2}. \tag{11}$$

Here, the velocity vectors $\mathbf{v}_\alpha$ and $\mathbf{v}_\beta$ of pedestrians α and β are included allowing to take into account the step size of pedestrians.

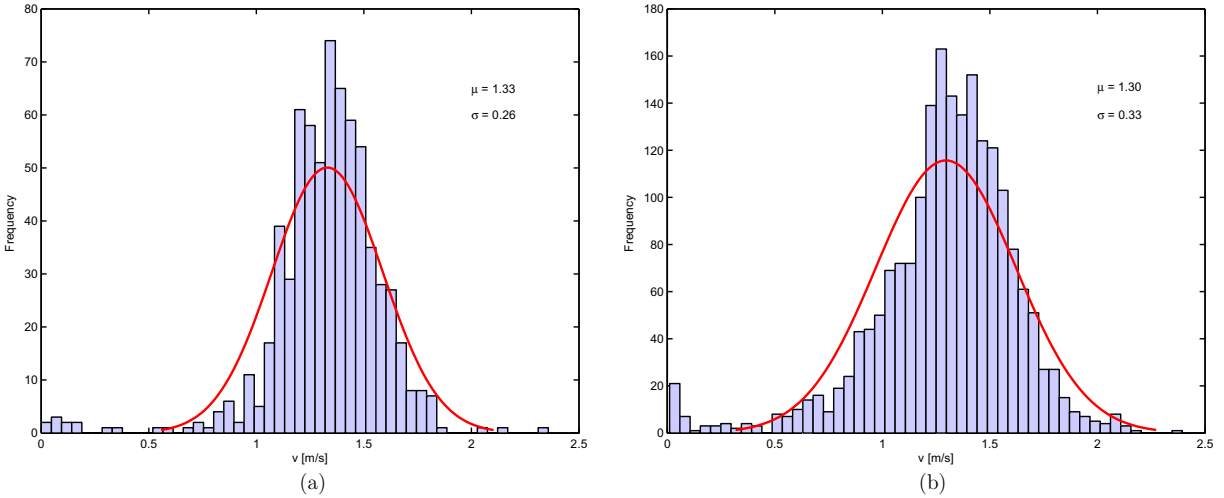


Fig. 11. Walking speed distribution from (a) experiment 1 and (b) experiment 2.

Model C: The third model is an implementation of Rudloff et al. (2011) in which the repulsive force is split into one force directed in the opposite of the walking direction, i.e. the *deceleration force*, and another one perpendicular to it, i.e. the *evasive force*. Here, the repulsive force is given as

$$\mathbf{f}^C_{\alpha\beta} = \mathbf{n}_\alpha \underbrace{a_n e^{\frac{-b_n \theta^2_{\alpha\beta}}{v_{\text{rel}}} - c_n \|\mathbf{d}_{\alpha\beta}\|}}_{\text{deceleration force}} + \mathbf{p}_\alpha \underbrace{a_p e^{\frac{-b_p |\theta_{\alpha\beta}|}{v_{\text{rel}}} - c_p \|\mathbf{d}_{\alpha\beta}\|}}_{\text{evasive force}}, \tag{12}$$

where $\mathbf{n}_\alpha$ is the direction of movement of pedestrian α and $\mathbf{p}_\alpha$ its perpendicular vector directing away from pedestrian β. Furthermore, $\theta_{\alpha\beta}$ is the angle between $\mathbf{n}_\alpha$ and $\mathbf{d}_{\alpha\beta}$ and v_{rel} denotes the relative velocity between pedestrians α and β.
We denote the implementations of the three above described repulsive formulations of the Social Force model as $\mathbf{f}^A_{\alpha\beta}$, $\mathbf{f}^B_{\alpha\beta}$ and $\mathbf{f}^C_{\alpha\beta}$. Note that the repulsive force from static obstacles $\mathbf{f}_{\alpha i}$ is modeled by using the same functional form as given by the repulsive force from pedestrians. Here, the point of an obstacle i closest to pedestrian α replaces the position β and $\mathbf{v}_i$ is set to zero. Furthermore, we take into account that pedestrians have a higher response to other pedestrians in front of them by including an anisotropic behavior, as described in Helbing and Johansson (2009), into the first two formulations.

4.3. Model calibration

The process of model calibration involves the identification of parameter values which produce realistic pedestrian behavior in the simulation results. We estimated values for the different parameters in the three described model approaches $\mathbf{f}^A_{\alpha\beta}$, $\mathbf{f}^B_{\alpha\beta}$ and $\mathbf{f}^C_{\alpha\beta}$ based on our empirical data set from the walking experiments. The trajectory data were divided into a non-overlapping calibration and validation data set (validation is described in Section 4.4) as shown in Table 3.

The literature describes different techniques for calibrating microscopic simulation models: one way is to estimate parameter values directly from the trajectory data by extracting pedestrian's acceleration (Hoogendoorn and Daamen, 2006). However, as shown in Rudloff et al. (2011) this method has several drawbacks, even with small errors in the trajectories. For instance, using the acceleration instead of the spatial position introduces significant noise due to the second derivative. Furthermore, this might lead to error-in-variables problems and parameter estimates possibly result in a bias towards zero.

Our calibration uses a simulation approach inspired by Johansson et al. (2007), where each pedestrian is simulated separately while keeping the remaining pedestrians on their observed trajectory. Each simulation run is performed according to the following procedure: the position and the desired goal for a simulated pedestrian α are extracted from the start point at time t^{in}_α and the end point at t^{out}_α of the associated observed trajectory $\mathcal{T}_\alpha$. The desired velocity v^0_α of pedestrian α is defined as the 90th percentile of the observed velocities. The magnitude of the current velocity vector $\mathbf{v}_\alpha$ is set equal to v^0_α, and it directs towards the pedestrian's desired goal. Pedestrian α is simulated for $M_\alpha = |\mathcal{T}_\alpha|$ timesteps during time t, with $t^{in}_\alpha \leqslant t \leqslant t^{out}_\alpha$, where both bounds are again derived from the observed trajectory. Note that we set for each simulated pedestrian $r = 0.2$ m and $\tau = 0.5$ s. It is left for future research to extract the actual radius of a pedestrian from the measured Kinect data.

After having simulated a set of N pedestrians from the calibration data set with the above procedure, the similarity measure s for testing the fit of our simulated trajectories is computed as

$$s = \frac{1}{N}\sum_{\alpha=1}^{N}\left(\frac{d(\alpha)}{t^{out}_\alpha - t^{in}_\alpha} + g(\alpha)\right). \tag{13}$$

For a pedestrian α, the mean Euclidean distance

$$d(\alpha) = d(\mathcal{T}_\alpha, \mathcal{T}^S_\alpha) = \frac{1}{M_\alpha}\sum_{i=1}^{M_\alpha}\left\|\mathbf{x}_{\alpha_i} - \mathbf{x}^S_{\alpha_i}\right\| \tag{14}$$

provides the dissimilarity between positions $\mathbf{x}_{\alpha_i} = [\mathbf{t}_{\alpha_i}, \mathbf{x}_{\alpha_i}, \mathbf{y}_{\alpha_i}]^T$ of the observed trajectory $\mathcal{T}_\alpha$ and positions $\mathbf{x}^S_{\alpha_i} = [\mathbf{t}^S_{\alpha_i}, \mathbf{x}^S_{\alpha_i}, \mathbf{y}^S_{\alpha_i}]^T$ of the simulated trajectory $\mathcal{T}^S_\alpha$. Furthermore, the length of trajectories is defined by $|\mathcal{T}_\alpha| = |\mathcal{T}^S_\alpha| = M_\alpha$. Since none of the used models explicitly restricts overlapping between pedestrians, an overlap penalty is added denoted by

Table 3
Partitioning of the trajectory data set for model calibration and validation.

	Number of trajectories	
	Experiment 1	Experiment 2
Calibration Set	516	1346
Validation Set	169	643
Total	685	1989

$$g(\alpha) = \frac{1}{N-1}\sum_{\beta\neq\alpha}\max_{t}\left(0, \frac{1}{\|\mathbf{d}_{\alpha\beta}(t)\|} - \frac{1}{r_\alpha + r_\beta}\right). \tag{15}$$

Model parameter values are estimated by applying an optimization algorithm to find the best possible fit by minimizing the objective function (13). We use a genetic algorithm which does not suffer from a starting value problem to find the neighborhood of the global minimum of (13). The estimated parameter values obtained by the genetic algorithm are then used as initial values for the Nelder-Mead algorithm (see Lagarias et al., 1998) to refine the result. This hybrid approach allows finding the global minimum while being numerically efficient.

4.4. Validation results

The results for the parameter fit of the individual models are provided in Table 4 as s_{cal} for the calibration data set and s_{val} for the validation data set. The best possible value for (13) is $s = 0$. For both experiments, the best fit of the objective function with the compared modeling approaches could be achieved using the repulsive formulation from $\mathbf{f}^C_{\alpha\beta}$ defined in (12).

By applying the three Social Force models on only a small subset of our validation data set, their basic ability of representing pedestrian behavior can be evaluated in a qualitative manner. Fig. 12 shows the results of a simulation run with 19 pedestrians in the setting of experiment 1: the simulation results of the circular force formulation from $\mathbf{f}^A_{\alpha\beta}$ in Fig. 12a indicate that simulated pedestrians evade relatively late with a strong deceleration caused by the static person in the center. In order to avoid running into the obstacle, some pedestrians even move slightly backward from the obstacle. This collision avoidance behavior differs significantly from the observed trajectories. As illustrated in Fig. 12b, the walking behavior from the simulations with $\mathbf{f}^B_{\alpha\beta}$ is less abrupt as a result of the included velocity dependence. However, pedestrian deceleration is again unrealistically strong when individuals directly approach the static obstacle. From a qualitative point of view,

Table 4
Fit of the parameter values for three different Social Force formulations based on calibration and validation data set.

	Experiment 1			Experiment 2		
	$\mathbf{f}^A_{\alpha\beta}$	$\mathbf{f}^B_{\alpha\beta}$	$\mathbf{f}^C_{\alpha\beta}$	$\mathbf{f}^A_{\alpha\beta}$	$\mathbf{f}^B_{\alpha\beta}$	$\mathbf{f}^C_{\alpha\beta}$
s_{cal}	0.1256	0.1105	0.0843	0.1721	0.1717	0.1630
s_{val}	0.0963	0.0823	0.0671	0.0967	0.0912	0.0883

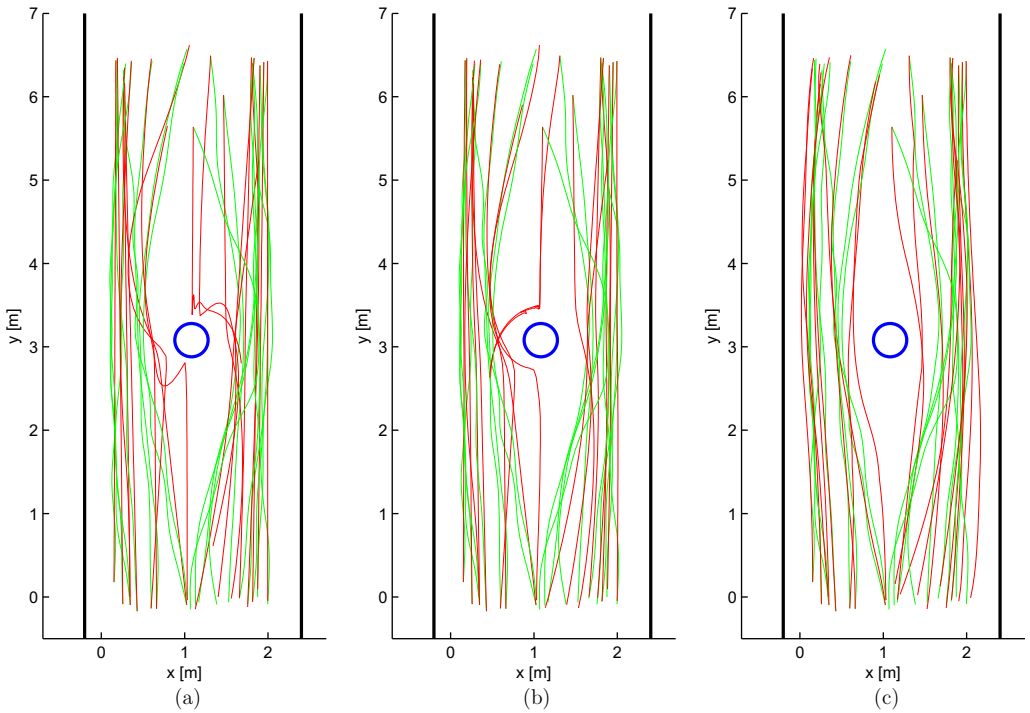


Fig. 12. Validation results of different Social Force models showing observed (green) and simulated trajectories (red) using (a) $\mathbf{f}^A_{\alpha\beta}$, (b) $\mathbf{f}^B_{\alpha\beta}$ and (c) $\mathbf{f}^C_{\alpha\beta}$ as repulsive force. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

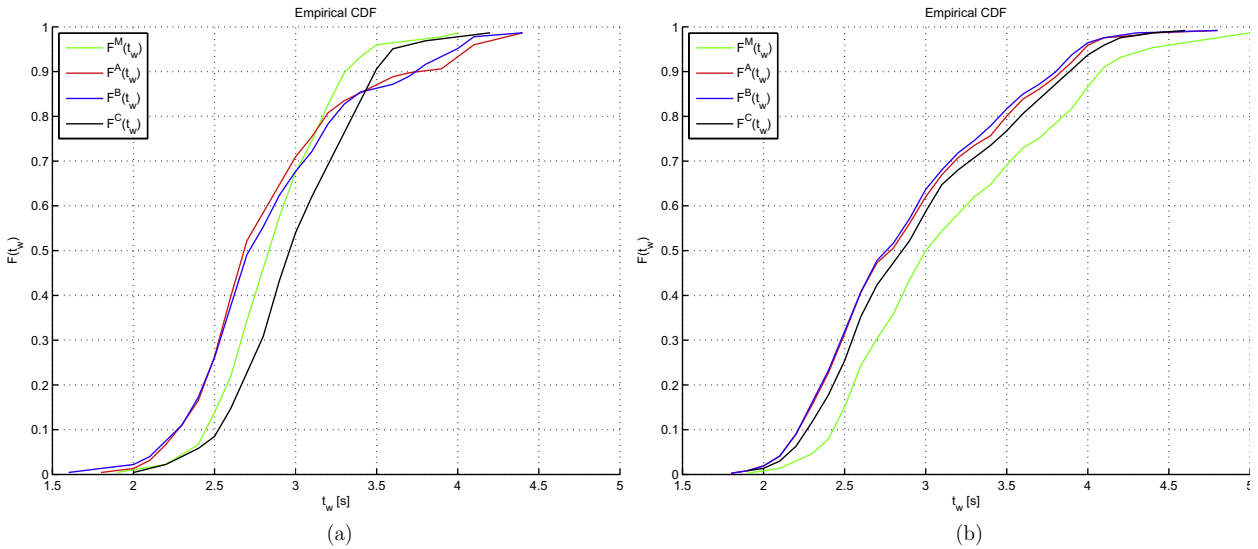


Fig. 13. Measured and simulated walking time distributions from (a) experiment 1 and (b) experiment 2.

simulation results obtained by using $\mathbf{f}^C_{\alpha\beta}$ exhibit the best results in our comparison (see Fig. 12c). Separating the forces into a deceleration and an evasive component results in individual trajectories which match very well with the observations.

For capacity estimations in infrastructures the walking times of pedestrians are of particular importance. Accordingly, pedestrian simulation models need to be able to reproduce realistic walking times even if they are not specifically calibrated for this purpose. Since the models in this work were calibrated using the similarity of trajectories as the objective function, we also want to evaluate their ability to correctly predict the walking time distribution based on our validation data set. Fig. 13 shows the cumulative distribution functions F^M of walking times t_w derived from measured trajectories, and the cumulative distribution functions F^A, F^B, F^C of the simulated trajectories provided by $\mathbf{f}^A_{\alpha\beta}, \mathbf{f}^B_{\alpha\beta}, \mathbf{f}^C_{\alpha\beta}$ respectively. The results for experiment 1 (see Fig. 13a) demonstrate that the cumulative walking time distribution F^A for the circular formulation and F^B for the elliptical formulation for the repulsive force in the Social Force model significantly deviate from the measured walking time distribution F^M. The formulation $\mathbf{f}^C_{\alpha\beta}$ provides a good replication of the measured walking time distribution F^M. In order to support this finding, we used a two-sample Kolmogorov–Smirnov test (see Massey, 1951) to compare each walking time distribution from the simulations with the measured distribution F^M. For a significance level of 0.05, we can reject the null hypothesis that F^A and F^M as well as F^B and F^M are from the same continuous distribution. However, the null hypothesis holds when comparing F^C and F^M.

5. Conclusion

In this work we have developed algorithms to use the Microsoft Kinect – basically a camera that also records 3-dimensional information in the form of a depth image – for automatic data collection of pedestrian movement from an elevated view. We have shown that the use of the Kinect allows the automated capture of human motion trajectories with high accuracy and without privacy issues. We applied our tracking algorithm to collect an extensive data set in the MIT's Infinite Corridor for calibrating and comparing three variations of the Social Force model.

Our approach groups depth information from a single Kinect in the world coordinate system into individual pedestrians based on hierarchical clustering. These detections are tracked over time throughout the sensing areas of multiple Kinects to obtain individual trajectories in larger space.

Evaluating the detection performance with two manually annotated ground truth data sets shows a Pedestrian Detection Rate of 94% and 96%, respectively. The position error for all correctly tracked objects is quantified as Multiple Object Tracking Precision and reveals relatively small values of around 4 cm. Compared to other scientific work such as Heath and Guibas (2008), where MOTP was in the range between 16 and 19 cm our approach shows significantly higher accuracy.

In conclusion, our tracking approach is capable of delivering trajectories with an accuracy which we consider sufficient for calibrating microscopic pedestrian simulation models. In the future our approach could be extended in order to also estimate the orientation of body parts, i.e. head and shoulder pose. This would allow us to gain more data on how humans perceive and interact with their environment which is particularly useful for evaluating visual information systems, such as guidance systems or lights.

By applying our tracking approach in two walking experiments performed under real world conditions in the MIT's Infinite Corridor, we gathered a total of 2674 trajectories. We compared three variations of the Social Force model by calibrating

them with our trajectory data. The validation results revealed that collision avoidance behavior in the Social Force model can be improved by including the relative velocity between individuals. Furthermore, dividing the repulsive force into a deceleration and an evasion part delivered the best quantitative and qualitative results out of the investigated models. However, dividing the repulsive force leads to a larger number of parameters, which makes the calibration process itself more complex and computationally expensive.

For future work we will increase our data set by obtaining trajectories under additional settings. Hence, other movement phenomena (e.g. crossing, uni-directional movements, turning of corners, etc.) can be observed which will allow for validating microscopic pedestrian simulation models on the operational level completely. This will also allow us to further investigate the transferability of different models to different scenarios. We also intend to automate the sensor calibration process. For example, the accuracy of an approach performing registration on robust 3D point sets such as described in Mavrinac et al. (2010) could be compared to the manual setup. Furthermore, including Kalman filters for pedestrian tracking might be useful for real-time applications performing more complex object detection algorithms which would not allow processing at the full Kinect framerate.

Going forward we believe that the adoption of the Kinect could be extremely useful for the development and calibration of pedestrian models – but also as a tool to better understand human crowd behavior and hence provide invaluable input to the design of all those spaces that need to respond to it – starting from our cities.

Acknowledgments

We would like to thank Jim Harrington, of the MIT School of Architecture and Planning, and Christopher B. Dewart, of the MIT Department of Architecture, for their support in the installation of equipment in the MIT's Infinite Corridor. The authors would also like to thank David Lee for his research assistance in designing and running the experiments. Support is gratefully acknowledged from the MIT SMART program, the MIT CCES program, Audi-Volkswagen, BBVA, Ericsson, Ferrovial, GE and all the members of the Senseable City Consortium.

References

Antonini, G., Bierlaire, M., Weber, M., 2006. Discrete choice models of pedestrian walking behavior. Transp. Res. Part B: Methodol. 40, 667–687.
Baltieri, D., Vezzani, R., Cucchiara, R., 2011. 3DPes: 3D people dataset for surveillance and forensics. In: Proceedings 1st International ACM Workshop on Multimedia access to 3D Human Objects, Scottsdale, Arizona, USA. pp. 59–64.
Berclaz, J., Fleuret, F., Turetken, E., Fua, P., 2011. Multiple object tracking using K-shortest paths optimization. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 33, 1806–1819.
Bernardin, K., Stiefelhagen, R., 2008. Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP J. Image Process. 2008, 1:1–1:10.
Berrou, J., Beecham, J., Quaglia, P., Kagarlis, M., Gerodimos, A., 2007. Calibration and validation of the legion simulation model using empirical data. In: Waldau, N., Gattermann, P., Knoflacher, H., Schreckenberg, M. (Eds.), Proceedings of the Conference on Pedestrian and Evacuation Dynamics (PED 2005). Springer, Berlin, Heidelberg, pp. 167–181.
Bob Fisher, 2010. Edinburgh Informatics Forum Pedestrian Database. http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/ (accessed October 2013).
Boltes, M., Seyfried, A., Steffen, B., Schadschneider, A., 2010. Automatic extraction of pedestrian trajectories from video recordings. In: Klingsch, W.W.F., Rogsch, C., Schadschneider, A., Schreckenberg, M. (Eds.), Proceedings of the Conference on Pedestrian and Evacuation Dynamics (PED 2008). Springer, Berlin, Heidelberg, pp. 43–54.
Burstedde, C., Klauck, K., Schadschneider, A., Zittartz, J., 2001. Simulation of pedestrian dynamics using a 2-dimensional cellular automaton. Physica A 295, 507–525.
Calderara, S., Heinemann, U., Prati, A., Cucchiara, R., Tishby, N., 2011. Detecting anomalies in people's trajectories using spectral graph analysis. Comput. Vision Image Understanding 115, 1099–1111.
Choi, W., Pantofaru, C., Savarese, S., 2013. A general framework for tracking multiple people from a moving camera. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 35, 1577–1591.
Daamen, W., Hoogendoorn, S.P., 2006. Free speed distributions for pedestrian traffic. In: Proceedings of the Transportation Research Board 85st Annual Meeting (TRB 2006), Washington D.C., USA. pp. 1–13.
Daamen, W., Hoogendoorn, S.P., 2012. Calibration of pedestrian simulation model for emergency doors for different pedestrian types. In: Proceedings of the Transportation Research Board 91st Annual Meeting (TRB 2012), Washington D.C., USA.
Davidich, M., Köster, G., 2012. Towards automatic and robust adjustment of human behavioral parameters in a pedestrian stream model to measured data. Safety Sci. 50, 1253–1260.
de Boor, C., 2001. A Practical Guide to Splines, second ed. Springer-Verlag, New York.
Duda, R., Hart, P., Stork, D., 2001. Pattern Classification. Wiley.
Duives, D.C., Daamen, W., Hoogendoorn, S.P., 2013. State-of-the-art crowd motion simulation models. Transportation Research Part C: Emerging Technologies.
Eiter, T., Mannila, H., 1994. Computing Discrete Fréchet Distance. Technical Report CD-TR 94/64. Vienna University of Technology.
Eshel, R., Moses, Y., 2010. Tracking in a dense crowd using multiple cameras. Int. J. Comput. Vision 88, 129–143.
Fleuret, F., Berclaz, J., Lengange, R., Fua, P., 2008. Multicamera people tracking with a probabilistic occupancy map. IEEE Trans. Pattern Anal. Mach. Intell. 30, 267–282.
Forsyth, D.A., Ponce, J., 2002. Computer Vision: A Modern Approach, 1 ed. Prentice Hall.
Frati, V., Prattichizzo, D., 2011. Using kinect for hand tracking and rendering in wearable haptics. In: Proceedings of the IEEE World Haptics Conference (WHC 2011), pp. 317 –321.
Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A., 2011. Efficient regression of general-activity human poses from depth images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2011). IEEE Computer Society, Los Alamitos, CA, USA, pp. 415–422.
Guo, R.Y., Wong, S., Huang, H.J., Zhang, P., Lam, W.H., 2010. A microscopic pedestrian-simulation model and its application to intersecting flows. Physica A: Stat. Mech. Appl. 389, 515–526.
Heath, K., Guibas, L.J., 2008. Multi-Person tracking from sparse 3d trajectories in a camera sensor network. In: Proceedings of the Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2008). IEEE, pp. 1–9.
Helbing, D., Johansson, A., 2009. Pedestrian, crowd and evacuation dynamics. Encycl. Complex. Syst. Sci. 16, 6476–6495.
Helbing, D., Molnár, P., 1995. Social force model for pedestrian dynamics. Phys. Rev. E 51, 4282–4286.

Helbing, D., Molnár, P., 1997. Self-Organization Phenomena in Pedestrian Crowds. Self-Organiz. Complex Struct. Individual Collective Dyn., 569–577.
Hirasaki, E., Moore, S., Raphan, T., Cohen, B., 1999. Effects of walking velocity on vertical head and body movements during locomotion. Exp. Brain Res. 127, 117–130.
Hoogendoorn, S., Daamen, W., 2006. Microscopic parameter identification of pedestrian models and implications for pedestrian flow modelling. Transp. Res. Rec. 1982, 57–64.
Hoogendoorn, S.P., Bovy, P.H.L., 2004. Pedestrian route-choice and activity scheduling theory and models. Transp. Res. Part B: Methodol. 38, 169–190.
Hoogendoorn, S.P., Daamen, W., 2003. Extracting microscopic pedestrian characteristics from video data. In: Proceedings of the Transportation Research Board 82st Annual Meeting (TRB 2003), Washington D.C., USA. pp. 1–15.
Hoogendoorn, S.P., Daamen, W., 2005. Pedestrian behavior at bottlenecks. Transp. Sci. 39, 147–159.
Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A., 2011. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST 2011). ACM, New York, NY, USA, pp. 559–568.
Javed, O., Shafique, K., Rasheed, Z., Shah, M., 2007. Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views. Comput. Vision Image Understanding 109, 146–162.
Johansson, A., Helbing, D., 2010. Analysis of empirical trajectory data of pedestrians. In: Klingsch, W.W.F., Rogsch, C., Schadschneider, A., Schreckenberg, M. (Eds.), Proceedings of the Conference on Pedestrian and Evacuation Dynamics (PED 2008). Springer, Berlin, Heidelberg, pp. 203–214.
Johansson, A., Helbing, D., Shukla, P.K., 2007. Specification of the social force pedestrian model by evolutionary adjustment to video tracking data. Adv. Complex Syst. 10, 271–288.
Klein, W., Köster, G., Meister, A., 2010. Towards the calibration of pedestrian stream models. In: Proceedings of the 8th International Conference on Parallel Processing and Applied Mathematics: Part II. Springer-Verlag, Berlin, Heidelberg, pp. 521–528.
Konolige, K., Mihelich, P., 2012. Technical Description of Kinect Calibration. http://www.ros.org/wiki/kinect_calibration/technical. (accessed October 2013).
Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E., 1998. Convergence properties of the Nelder-mead simplex method in low dimensions. SIAM J. Optimiz. 9, 112–147.
Loy, C., Xiang, T., Gong, S., 2009. Multi-camera activity correlation analysis. In: Proceedings IEEE Intl. Conference on Computer Vision and Pattern Recognition (CVPR09), pp. 1988–1995.
Massey, F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46, 68–78.
Mavrinac, A., Chen, X., Tepe, K., 2010. An automatic calibration method for stereo-based 3D distributed smart camera networks. Comput. Vision Image Understanding 114, 952–962.
Microsoft Corp., 2012a. Kinect for Xbox 360. Redmond, WA, USA.
Microsoft Corp., 2012b. Microsoft Kinect for Windows SDK. http://www.microsoft.com/en-us/kinectforwindows/. (accessed October 2013).
Moussaïd, M., Helbing, D., Theraulaz, G., 2011. How simple rules determine pedestrian behavior and crowd disasters. Proc. Natl. Acad. Sci. 108, 6884–6888.
Moussaïd, M., Perozo, N., Garnier, S., Helbing, D., Theraulaz, G., 2010. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PLoS One 5, e10047.
Noonan, P.J., Cootes, T.F., Hallett, W.A., Hinz, R., 2011. The design and initial calibration of an optical tracking system using the microsoft kinect. In: Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC 2011), pp. 3614–3617.
Oh, S., Hoggs, A., Perero, A., Cuntoor, N., Chen, C., Lee, J., Mikherjee, S., Aggarwal, J., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., M., D., 2011. A Large-scale benchmark dataset for event recognition in surveillance video. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR2011). IEEE.
Pellegrini, S., Ess, A., Schindler, K., van Gool, L., 2009. You'll never walk alone: modeling social behavior for multi-target tracking. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2009). IEEE Computer Society, Los Alamitos, CA, USA, pp. 261–268.
PETS, 2009. PETS Dataset - Performance Evaluation of Tracking and Surveillance. http://www.cvg.rdg.ac.uk/PETS2009/. (accessed October 2013).
Pflugfelder, R., Bischof, H., 2010. Localization and trajectory reconstruction in surveillance cameras with non overlapping views. IEEE Trans. Pattern Anal. Mach. Intell. 34, 709–721.
Plaue, M., Chen, M., Bärwolff, G., Schwandt, H., 2011. Trajectory extraction and density analysis of intersecting pedestrian flows from video recordings. In: Proceedings of the 2011 ISPRS Conference on Photogrammetric Image Analysis (PIA 2011). Springer, Berlin, Heidelberg, pp. 285–296.
PTV Group, 2013. PTV Viswalk. http://vision-traffic.ptvgroup.com/en-us/products/ptv-viswalk/. (accessed October 2013).
Rudloff, C., Matyus, T., Seer, S., Bauer, D., 2011. Can walking behavior be predicted? An analysis of the calibration and fit of pedestrian models. Transp. Res. Rec. 2264, 101–109.
Savannah Simulations, 2013. SimWalk. http://www.simwalk.com/. (accessed October 2013).
Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-time human pose recognition in parts from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011). IEEE Computer Society, Los Alamitos, CA, USA, pp. 1297–1304.
Sochman, J., Hogg, D., 2011. Who knows who – inverting the social force model for finding groups. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE.
Stauffer, C., Grimson, W.E.L., 2000. Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 22, 747–757.
Sternig, S., Mauthner, T., Irschara, A., Roth, P., Bischof, H., 2011. Multi-camera Multi-object Tracking by Robust Hough-based Homography Projections, in: Proceedings of the Eleventh IEEE International Workshop on Visual Surveillance (ICCV Workshops), pp. 1689–1696.
United Nations, D.o.E., Affairs, S., 2012. World urbanization prospects: The 2011 revision highlights.
University of Cyprus Computer Graphics Lab, 2011. Crowds-by-Example data set. https://graphics.cs.ucy.ac.cy/research/downloads/crowd-data. (accessed October 2013).
Weiss, A., Hirshberg, D., Black, M., 2011. Home 3D body scans from noisy image and range data. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2011), Barcelona. pp. 1951–1958.