Automatic inference of geometric camera parameters and inter-camera topology in uncalibrated disjoint surveillance cameras

Richard J.M. den Hollander, Henri Bouma, Jan Baan, Pieter T. Eendebak, Jeroen H.C. van Rest

TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands

ABSTRACT

Person tracking across non-overlapping cameras and other types of video analytics benefit from spatial calibration information that allows an estimation of the distance between cameras and a relation between pixel coordinates and world coordinates within a camera. In a large environment with many cameras, or for frequent ad-hoc deployments of cameras, the cost of this calibration is high. This creates a barrier for the use of video analytics. Automating the calibration allows for a short configuration time, and the use of video analytics in a wider range of scenarios, including ad-hoc crisis situations and large scale surveillance systems. We show an autocalibration method entirely based on pedestrian detections in surveillance video in multiple non-overlapping cameras. In this paper, we show the two main components of automatic calibration. The first shows the intra-camera geometry estimation that leads to an estimate of the tilt angle, focal length and camera height, which is important for the conversion from pixels to meters and vice versa. The second component shows the inter-camera topology inference that leads to an estimate of the distance between cameras, which is important for spatio-temporal analysis of multi-camera tracking. This paper describes each of these methods and provides results on realistic video data.

Keywords: Autocalibration, pedestrian detection, tracking, intra-camera geometry, inter-camera topology

1. INTRODUCTION

Nowadays, surveillance cameras are used for the monitoring of persons at places like train stations, shopping malls and airports. Since the number of cameras has increased rapidly, automated video content analysis (VCA) would be desirable to assist camera operators, e.g. for finding suspects[1] and detecting certain events or specific behaviours.[2,3] One of the requirements for many VCA tools is the availability of a camera calibration, which can be used to relate pixel coordinates to world coordinates.[4,5] This will allow measurements of walking speed, or distance between objects in the scene. It is important for activity analysis in a single camera, like detection of persons running or loitering, see Fig. 1.

Some activities have a spatial extent that is larger than the coverage of a single camera view, and can only be concluded after observing a person in several camera views. This motivated the development of inter-camera tracking methods and inter-camera activity analysis. A person's trajectory that is determined by inter-camera tracking, as well as any detected single-camera activities, will enable the analysis of long term activity for a person. For reliable inter-camera tracking, a network of multiple non-overlapping cameras needs a joint calibration, so that the spatial relations between the cameras are known.

Both the calibration of individual cameras (referred to as intra-camera geometry) and the network of cameras (referred to as inter-camera topology) in a large environment come with high cost. This creates a barrier for the use of video analytics. Automating the calibration allows for a short configuration time, and the use of video analytics in a wider range of scenarios, including ad-hoc crisis situations and large scale surveillance systems. The intra-camera geometry as well as the inter-camera topology can be derived from the presence of persons in the camera views. Since there are usually persons present in surveillance scenarios, research has focused on

E-mail: richard.denhollander@tno.nl

R.J.M. den Hollander, H. Bouma, J. Baan, P.T. Eendebak, J.H.C. van Rest, "Automatic inference of geometric camera parameters and inter-camera topology in uncalibrated disjoint surveillance cameras," Proc. SPIE, Vol. 9652, (2015). http://dx.doi.org/10.1117/12.2194435

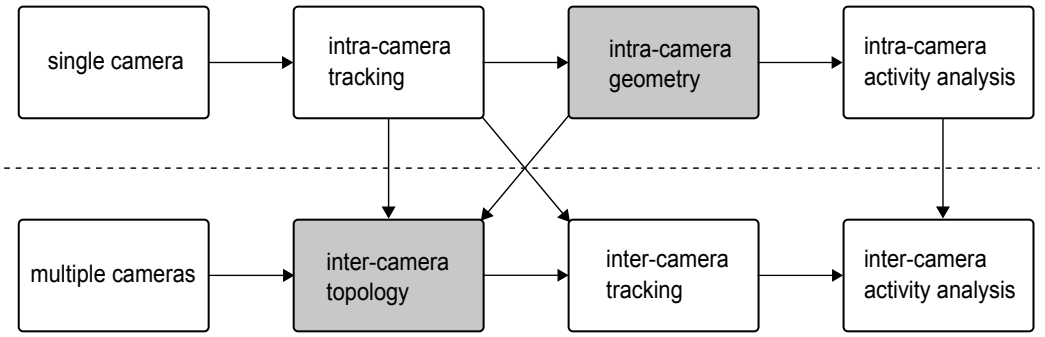


Figure 1. The successive processing steps for video analytics in single or multiple camera (network) configurations. Intra-camera tracking provides person (pixel) positions over time, which can be used for intra-camera geometry estimation. The intra-camera geometry will enable subsequent activity analysis for speed and distance related activities. In case of multiple cameras in a network, an inter-camera topology will improve inter-camera tracking and activity analysis across cameras. The computation of inter-camera topology uses the intra-camera geometries and single camera tracks. This paper focusses on automatic calibrations for the intra-camera geometry and the inter-camera topology (gray boxes).

methods for using them as calibration objects. For intra-camera geometry, head and foot locations of detected persons can be used for the estimation of the calibration parameters. The calibration parameters refer here to both intrinsic and extrinsic camera parameters. The intrinsic parameters like focal length combined with local extrinsic parameters like tilt angle and height are all needed for doing measurements in the scene. In the first part of this paper, we propose a new calibration method for these parameters that is based on the pixel heights of persons in the scene. It is a fully automatic method that collects the persons' measurements and produces the required camera parameters. In the second part of this paper, we present an inter-camera calibration method based on correlations between predicted and measured passing sequences of persons in camera views. This method enables the automatic determination of mutual camera distances and network topology. The proposed methods are tested on recordings from a camera network in a shopping mall.

The paper is organized as follows. In Section 2 the (single) intra-camera calibration approach is presented, where Section 2.2 describes the method and Section 2.3 presents experimental results. In Section 3 the inter-camera topology approach is presented, with a method section (Sec. 3.2) and a results section (Sec. 3.3). Finally, Section 4 contains our conclusions.

2. INTRA-CAMERA GEOMETRY

2.1 Background

Computing the intra-camera calibration requires several correspondences between image coordinates and their world positions, or, alternatively, building structures suitable for determining the vanishing line (horizon) and vertical vanishing point in the scene. Unfortunately, such elements are not always visible in the scene. Since there are usually persons present in surveillance scenarios, they have been proposed as alternative calibration object. Lv et al.[6] have explored this concept and extracted head and foot locations of detected persons for the estimation of the horizon line and vertical vanishing point. The same constraints are used by Krahnstoever and Mendonca[7] where they have added a statistical analysis of the errors, which allowed higher levels of noise and outliers among the data. Their work was extended in[8] with the incorporation of an additional constraint on the persons' walking speed, and Junejo and Foroosh[9] made an extension with an improved estimation process of the parameters. A RANSAC-based approach for finding the vertical vanishing point was proposed by Liu et al.,[10] where, in addition, statistics about the population's height were used in the verification of parameter estimates. Most of these approaches assume that the focal length is the only intrinsic parameter to be estimated, and that other intrinsics like aspect ratio, the principal point and skew have default values (see Hartley and Zisserman[11] for details on the camera intrinsics). Mohedano and Garcia[12] showed that it is not possible to compute all intrinsics using only horizon line and vertical vanishing point.

The vertical vanishing point in these methods is susceptible to small errors in the extracted head and feet positions of the detected persons, since the persons in the scene are only small line segments for which the common intersection point must be found. Instead, we propose a method based on a different type of measurement, namely the pixel height of persons in the scene. This also uses knowledge of people's head and feet locations, but only requires the height difference and no directional information. We will derive the calibration parameters from the pixel height distribution (see Fig. 2), which contains the pixel height of persons for different feet positions in the image. The pixel height distribution can be derived automatically by using the bounding boxes from a pedestrian detector. The calibration process can therefore be integrated with pedestrian detection, which is a standard component for video analytics in many surveillance scenarios.

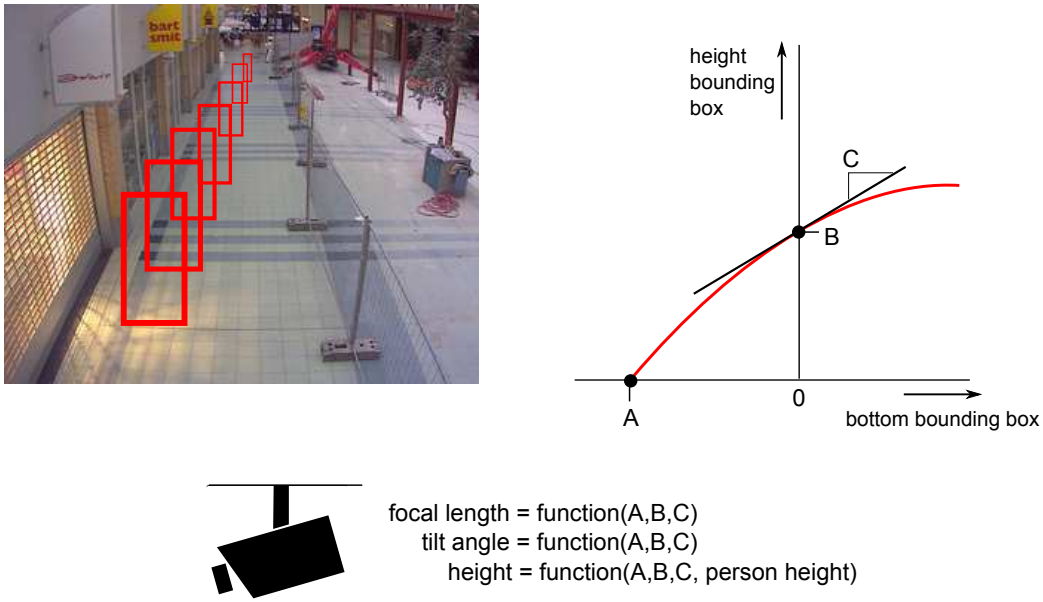


Figure 2. Overview of the intra-camera calibration method. The pixel heights of a person across the scene are measured by means of a pedestrian detector (upper left). The heights of the detections are plotted against the person's feet position (upper right). Three characteristics of this height distribution are measured: the horizon y-value (A), the person height at the image centre (B) and the height derivative at the image centre (C). Under the assumption of constant person height in the scene, these three measurements enable the computation of the camera's focal length, its tilt angle and its height. For the camera height computation, the person height in metres should also be known. The height distribution is obtained by averaging the detections of multiple persons. (The figure is best viewed in colour.)

2.2 Camera parameters from pedestrian heights

There are a few assumptions that we make about the camera, namely that it has zero skew, square pixels and the principal point at the image centre (which coincides with the origin of the image coordinate system). In this case it holds for the camera calibration matrix that $K = \mathrm{diag}(f, f, 1)$. The cameras in our experiments are *not* selected based on this criteria, so any deviation from our assumption will influence the results. Furthermore, we assume that the lens distortion is negligible or can be compensated for. Unlike the camera parameters that we will estimate and that can be subject to change during camera install or operation, the lens distortion only needs to be determined once and remains fixed afterwards. In our experiments we will correct all bounding-box coordinates for the lens distortion. Finally, the camera is assumed to have zero roll angle. In case the roll angle is non-zero there are other methods available to compensate for this (see e.g. Richardson et al.[13]), but this is currently not taken into account. Based on the foregoing assumptions, the only internal camera parameter that is estimated is the focal length, and the estimated external parameters are the camera tilt angle and height.

We have chosen the camera setup as shown in Fig. 3, where the origins of the world and camera coordinate frame coincide in point **C**. The camera is looking downward with a tilt angle $0 < \theta < 90°$, and is h_{cam} metres above the ground plane level. A person of height h_{person} is observed by the camera, and is projected onto the image plane between y-coordinates y_{feet} and y_{head}. We will denote the height of the person's bounding box by

$$\Delta y \equiv y_{\text{feet}} - y_{\text{head}} \ . \tag{1}$$

We further introduce the notation Δy_{pp} which is the bounding-box height at the principal point of the camera where $y_{\text{feet}} = 0$. The derivative of the bounding-box height with respect to the bottom position ($= y_{\text{feet}}$) at the principal point is denoted as $\partial \Delta y / \partial y_{\text{feet pp}}$.

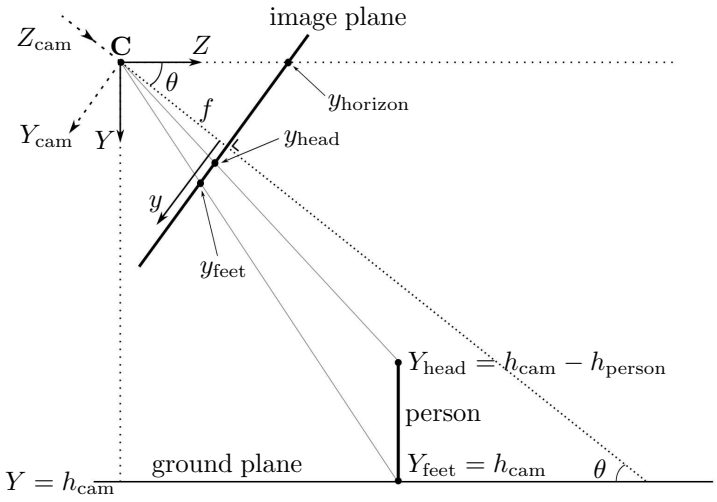


Figure 3. The camera setup for the calibration method. The camera is looking downward with a tilt angle θ, and is positioned h_{cam} metres above the ground plane level. The image plane contains the projected positions of a person's head and feet, and the projection of the horizon.

In order to compute the camera's focal length f, tilt angle θ and height h_{cam} we require three quantities, namely Δy_{pp}, $\partial \Delta y / \partial y_{\text{feet pp}}$ and y_{horizon}. The last quantity is the estimated y-coordinate of the horizon. Given these three quantities, it can be shown (see Appendix A) that the camera parameters follow from

$$f = \sqrt{\frac{\Delta y_{\text{pp}} y_{\text{horizon}}^2 + \Delta y_{\text{pp}}^2 y_{\text{horizon}}}{y_{\text{horizon}} \frac{\partial \Delta y}{\partial y_{\text{feet}}}_{\text{pp}} + \Delta y_{\text{pp}}}} \ , \tag{2}$$

$$\theta = \arctan\left(\frac{-y_{\text{horizon}}}{f}\right) \ , \tag{3}$$

$$h_{\text{cam}} = \frac{f h_{\text{person}} \sin\theta \cos\theta}{\Delta y_{\text{pp}}} + h_{\text{person}} \sin^2\theta \ . \tag{4}$$

Note that the person's height is only necessary for computation of the camera height; for the focal length and tilt angle it is only required that the person's height is constant.

The quantities Δy_{pp}, $\partial \Delta y / \partial y_{\text{feet pp}}$ and y_{horizon} are computed from the height distribution (see Fig. 2). The height distribution is approximated by a second degree polynomial of the pixel feet position. We therefore use a polynomial fit on the measured (average) height distribution in order to compute the desired quantities.

2.3 Experimental results

2.3.1 Setup

The experiments are performed on several surveillance cameras in a shopping mall (see left images in Fig. 4), where there are many passers-by during the course of a day. We have used 8 hours of video data in our experiments. Persons are detected in the raw images using the FPDW (fastest pedestrian detector in the west) detector from Dollar et al.,[14] and, subsequently detections with too small confidence values or with minimum bounding-box height are discarded. Next, the remaining detections are combined into tracks using a simple bounding-box overlap criterion. We perform additional track filtering (on track length and covered distance) in order to arrive at a reliable set of (track) bounding boxes. On average, around 10,000 tracks remain for each of the cameras, which contain between 700,000 and 1,100,000 bounding boxes for the camera.

We perform detection on the non-rectified imagery and correct the bounding boxes for the lens distortion using a ground-truth estimate. Without lens distortion, the bounding-box height of a person should be approximately equal across the width of the image. Thus, we can average all the bounding-box heights at a certain y-feet position in the image. It is still possible that detections errors will cause the height distribution to be a non-smooth curve, especially near the camera in the lower part of the image. Therefore, we fit the second-degree polynomial to the height distribution multiple times, and iteratively remove any large deviations in height on the right end of the distribution (i.e. closer to the camera).

The average person height h_{person} is chosen as 1.8 metres. For comparison, we also generate synthetic detections of persons exactly 1.8 metres in height. These bounding boxes are spread regularly across the image and are generated with the ground-truth camera calibration. They simulate the case of a perfect pedestrian detection algorithm with accurate height values, and give an upper bound on the accuracy that is attainable. Note that synthetic data will not yield results equal to the ground truth, since the cameras' principal points are not exactly at the image centre, and the second degree polynomial is not a perfect fit for the height distribution.

2.3.2 Results

The surveillance cameras used in the experiments are shown in Fig. 4 on the left side. The result of pedestrian detection and the iterative selection process are shown in the middle graphs. The green height values have been removed from the original distribution and the blue values remain. The red values correspond to the synthetically generated data. We see especially for camera 1 and to some degree for camera 3 that there is a deviation between synthetic and real data close to the camera. This is mainly caused by the relatively larger tilt angles of these cameras, so that pedestrians' extended legs while walking to/from the camera cause the detection bounding boxes to be oversized. We can only partly compensate for this effect by the selection process. Although the selection may as well exclude some correct data (see camera 2), there is still sufficient data at the principal point position for the (derivative of) height.

The camera parameters are estimated with the proposed method on both the real and synthetic detections, see Table 1 for the results. As can be seen, the use of synthetic data allows accurate parameter estimates with deviations smaller than 4% in focal length, 2.8 degrees in tilt angle and 17 cm in height. For the real detections, we see that both cameras 2 and 4 show results that are close to the ground-truth/synthetic values, and that cameras 1 and 3 have larger deviations for the real detections. A relevant question is what the influence of these parameter estimates is on the measurement of quantities in the scene. For this purpose, we have used a series of 50 synthetic detections that are positioned along the camera's principal axis up to 50 metres away. The detections are all 1.8 metres in height and spaced 1.0 metre apart. The projections in the camera images are combined with the estimated parameters to retrieve the height and spacing values. In Fig. 4 the graphs at the right show the results for both measurements (measure 1=height, measure 2=spacing). As expected, the estimated parameters for cameras 2 and 4 produce accurate measurements up to 50 metres away, with errors smaller than 4% for the heights and 10% for the distances. Although the parameter values showed significant deviations, camera 3 shows good height accuracy with errors smaller than 6.5%, and a somewhat smaller distance accuracy with errors up to 20% close to the camera. It turns out that the combination of parameters for camera 3 was estimated in such a way (too small tilt angle combined with too large focal length) that the resulting measurement errors are small. The parameter estimates for camera 1 produce measurement errors up to 16% and 50% for the heights and distances on 50 metres, respectively. On 30 metres distance, however, the height error for camera 1 is still

below 10%. As mentioned before, we witnessed oversized bounding boxes due to extended legs for this camera. This problem could be addressed with a method similar to the one developed by Lv et al.,[6] who tried to detect leg-crossing phases in the human walking pattern. Such an approach may bring the calibration accuracy on the same level as the other cameras.

	focal length f (pixels)	tilt angle θ (degrees)	height h_{cam} (metres)
camera 1			
ground truth	849.3	36.9	4.2
synthetic data	882.7	35.5	4.4
real data	1023.7	30.8	4.4
camera 2			
ground truth	1211.7	26.8	4.2
synthetic data	1185.3	25.1	4.2
real data	1186.9	25.0	4.3
camera 3			
ground truth	867.1	36.1	4.2
synthetic data	884.7	33.2	4.4
real data	1158.7	27.3	5.0
camera 4			
ground truth	1129.3	27.3	3.9
synthetic data	1135.3	26.4	3.9
real data	1079.4	27.3	3.8

Table 1. The results of parameter estimation for the cameras in Fig. 4. The first line for each camera shows the ground-truth calibration parameters. The second line shows the parameters estimated using synthetic detections, and the third line shows the estimation results based on real detections.

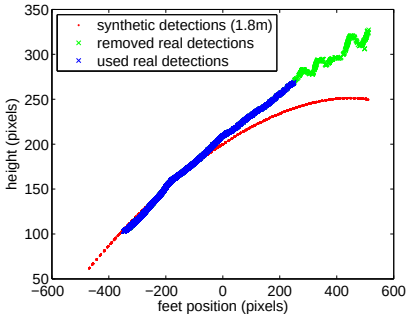
3. INTER-CAMERA TOPOLOGY

3.1 Background

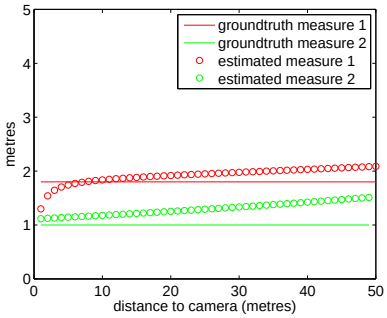
The inference of camera topology is an active research field[15–24] with comprehensive surveys[4,5] and many recent contributions.[25–30] In most camera networks the cameras will generally have non-overlapping views, so that they can not (all) be calibrated by common scene properties. Furthermore, a full site modelling for determining the camera positions is usually prohibitive.[5] Instead, the spatial relations between cameras can be found by analysing the person flow in the different camera views. In this context, the topology inference can be based on correspondence or correspondence-free methods.[5] Correspondence methods aim at explicit recognition of individual persons in order to link two camera views. Such methods can use different types of recognition, e.g. based on clothing appearance (Kuo and Huang[21]) or faces (Zou et al.[17]). In some cases the search for correspondences can be difficult, and this motivated the development of correspondence-free methods. These methods rely on the statistics of person arrivals in camera views, where a statistical relation between person flows indicates a spatial connection. Makris et al.[15] considered the cross-correlation of departure and arrival sequences of persons in automatically learned entry/exit zones. This resulted in estimates for the transition times between camera nodes in the network. Tieu et al.[16] extended this idea and introduced the use of mutual information for deriving statistical dependence between departure and arrival sequences. They showed that mutual information was better suited for multi-modal transition times, e.g. originating from a mix of pedestrians and vehicles. Wang et al.[22] have used a classification model for trajectories which clusters trajectories based on the activities performed.

We propose a new method inspired by the work of Makris et al.,[15] that differs in several aspects compared to their correlation method. First, the departure sequence is not directly correlated with the arrival sequence, but every person's departure from a camera entry/exit point is extrapolated based on its walking speed. In this way we can compensate for different walking speeds, which would otherwise be assumed constant. Second, all persons crossing a certain entry/exit point are used in the correlation, i.e. not only persons departing the camera

(a) Camera 1

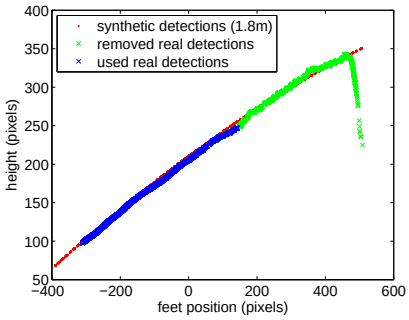


(b) Camera 1 detections

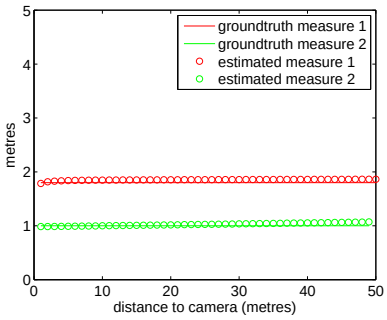


(c) Camera 1 scene measurements

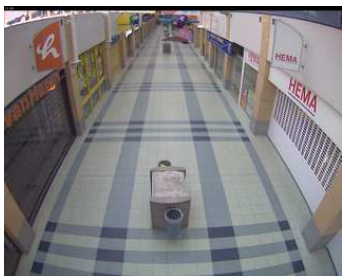
(d) Camera 2



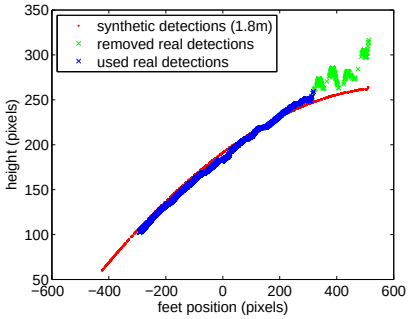
(e) Camera 2 detections



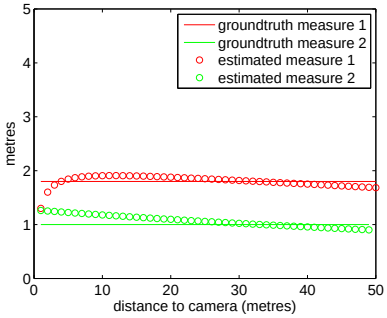
(f) Camera 2 scene measurements



(g) Camera 3

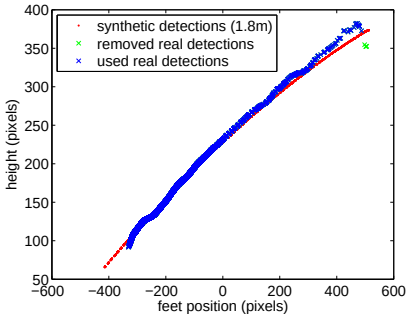


(h) Camera 3 detections

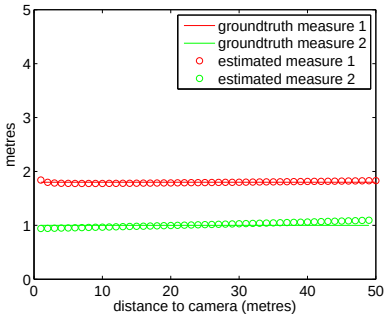


(i) Camera 3 scene measurements

(j) Camera 4



(k) Camera 4 detections



(l) Camera 4 scene measurements

Figure 4. *Left*: Camera images. *Middle*: Synthetic detections (red) are compared with real detections (green and blue) and the selected subset (blue). *Right*: The estimated camera parameters for the real detections are used to estimate the heights (measure 1, synth.=1.8m) and ground plane separations (measure 2, synth.=1.0m). (Best viewed in colour.)

view but also those arriving in view. This set of departures and arrivals combined, extrapolated with walking speeds, is correlated to the measured passing sequence in another camera view's entry/exit point. Furthermore, the distance estimates between camera nodes in the network allow us to reconstruct a camera floorplan of the environment. We emphasize that this reconstruction is entirely based on person tracks without any explicit site modelling.

3.2 Topology from entry/exit point passing sequences

The proposed inter-camera topology method consists of several automated steps, outlined below:

- determine (combined) entry/exit regions in each camera view
- extrapolate the position of all persons passing a region's boundary based on their walking speed
- correlate the extrapolated passing sequences for region A in one camera with the measured passing sequence for region B in another camera
- determine distances (and possibly time offsets) between the cameras by finding the maximum correlation scores
- build a graph from the distances by multi-dimensional scaling

The first step is the determination of regions in a camera view where persons are entering or exiting the scene. The region in the scene where there is a directed flow of people is potentially connected to another camera's entry/exit point, therefore establishing a topological relation. In order to find the flow regions, we first determine the perimeter of all person tracks in the camera view. Then we construct a histogram over this perimeter to count the number of tracks (both entering and exiting) that cross the perimeter when extrapolated, see Fig. 5. By means of thresholding we find a set of extended flow regions on the perimeter. In the example shown there are 4 regions of dense person flow; 2 near shop entries and 2 at both ends of the hallway.

The next step is the extrapolation of tracks crossing a camera region. In other words, the walking speed of a person is used to predict his/her position for different time lapses since the moment of crossing. This will result in a passing sequence of crossings for various distances, that could match the measured passing sequence of another region for the correct distance value, see Fig. 6. We will determine the distance by correlating the measured passing sequence with differently predicted passing sequences. In case the camera recordings are not synchronized, there is an additional time offset between the sequences. It is possible to simultaneously estimate any time offset together with the distance, see Fig. 7 for an illustration. The time offset is merely a shift of the passing sequence, and does not influence the mutual positions of the predicted positions. The correlation result will typically yield a peak at a certain distance and time offset value. When the camera regions are far apart, the number of persons that follows the route between the regions is small. Yet, even a small percentage of persons on that route will still produce a peak in the correlation results, see e.g. the correlation for regions 3 and 28. Note that the correlation is a linear relation between the distance and time offset. Due to the fact that the camera regions contain both entering and exiting tracks, we can see two linear patterns in the plot which intersect at the optimal distance/offset combination (where also the maximum correlation is observed). The bi-directional tracks are therefore especially important when there is no time synchronization and walking speeds are very similar. In this example, the time offsets are zero.

The correlation of the predicted and measured passing sequences will not be very precise due to small estimation errors in the persons' speed, and the presence of speed variations when the distance between cameras is large. Therefore, we first smooth the measured passing sequences so that small inaccuracies in predicted positions can be compensated for (see Fig. 6).

The pairwise distance relations between cameras are already useful for inter-camera tracking, since the transition time (for person re-appearance) in a camera pair is then known. We can, however, extend the pairwise distance relations to a topological network and reconstruct the camera floorplan. This enables the estimation of transition time between cameras at a large distance. For the purpose of topology computation, we select the highest correlations between all pairs of camera regions, and apply multi-dimensional scaling[31] on the pairwise

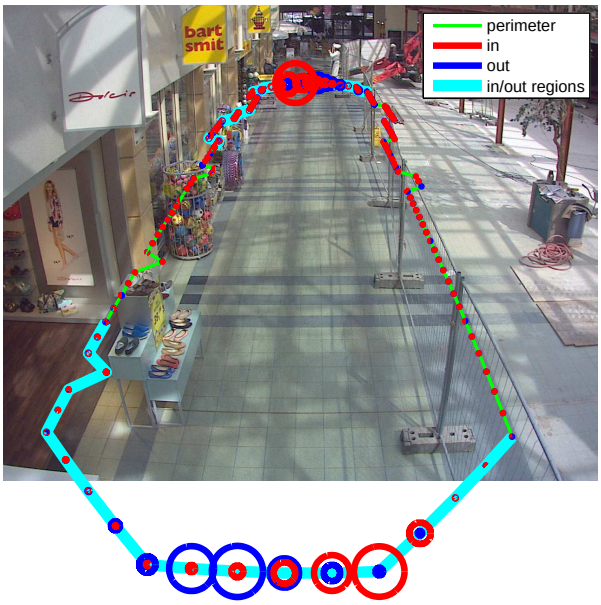


Figure 5. The perimeter of a camera view with the entrance (red) and exit (blue) directions overlaid. The major person flow directions are combined into regions; here 4 different entry/exit regions have been found. (The figure is best viewed in colour.)

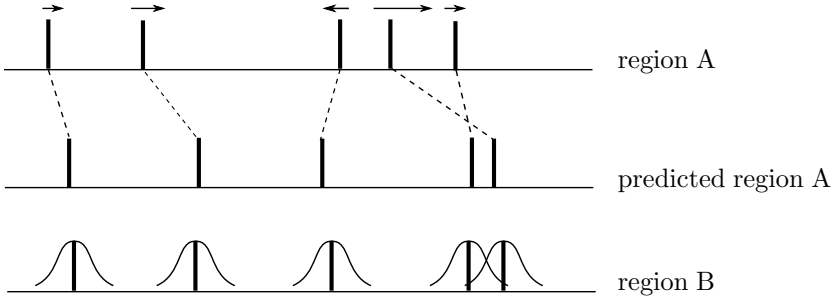
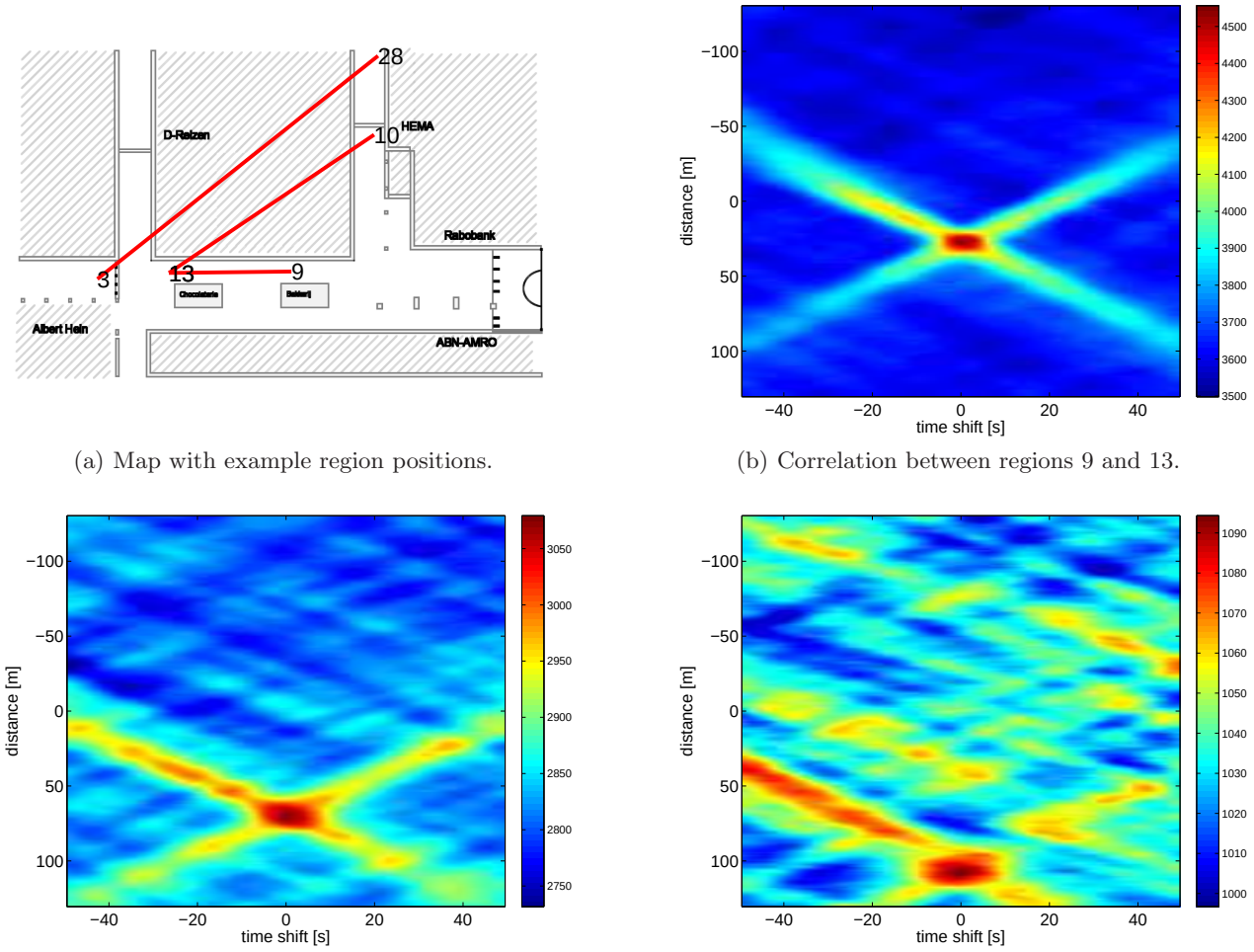


Figure 6. *Top line:* A sequence of persons that pass region A with different walking speeds (arrows) and in different directions. *Middle line:* A prediction of this sequence is made for a certain chosen distance value. *Bottom line:* The measured passing sequence of persons in another region B, that corresponds to region A for the chosen distance. The sequence is smoothed in order to be robust against small inaccuracies.

distances. This produces a two-dimensional graph, with Euclidean distances between the nodes approximately equal to the pairwise distances. The correlation values have been used to weight the distance inputs in the scaling operation, in order to give more importance to better inputs.

3.3 Experimental results

We have used 14 surveillance cameras from the shopping mall, see Fig. 4 for example views. In principle, we could use the calibration procedure from Section 2 for obtaining the intra-camera calibrations. However, we decided to use a manual calibration procedure based on 3D scene elements, since not all cameras were suitable for automatic calibration. In particular, some camera views had nonzero roll angle. In order to have a uniform intra-camera calibration for all cameras, we decided to use the manual procedure here. Another reason is that the quality of the final topology estimate should not be adversely affected by any small deviations in the intra-camera calibrations. In total 4 hours of video footage was used for the generation of person detections and subsequent



(a) Map with example region positions.

(b) Correlation between regions 9 and 13.

(c) Correlation between regions 10 and 13.

(d) Correlation between regions 3 and 28.

Figure 7. Examples of correlations between regions from several camera positions. The distance between the cameras is increasing from the first until the third example. (The figures are best viewed in colour.)

tracking. The tracks were then used to determine the camera flow regions. The resulting person flow regions are shown in Fig. 8.

The resulting regions are not very accurate yet. For example, the large green region in the centre right should have been subdivided into two or three regions. A view of a large open area, where persons are crossing from all directions, is namely more difficult to subdivide into regions than a hallway view. Another cause is that the perimeter of all tracks in a camera view lies far away, and flow estimates become less reliable at a distance. In order to prevent the topology from being inaccurate because of just region quality, we have used a manual determination of the regions in the remainder of the experiments.

The predicted passing sequences from the regions are correlated with the measured passing sequences. In order to see how well the maximum correlations between camera regions indicate real relations, we have shown the correlations between all regions in a single plot, see Fig. 9. For each region pair, the estimated distance is plotted against the ground truth straight line distance between the regions. Although the straight line distance is an underestimate, it provides an indication of the correctness of the correlation results. In particular, the plot shows a clustering of high correlation values on the diagonal, indicating successful correspondences between camera regions. The resulting graph acquired with multi-dimensional scaling is shown in Fig. 10. It is very similar to the actual camera region floorplan; the hallway structure and relative region positions are reproduced

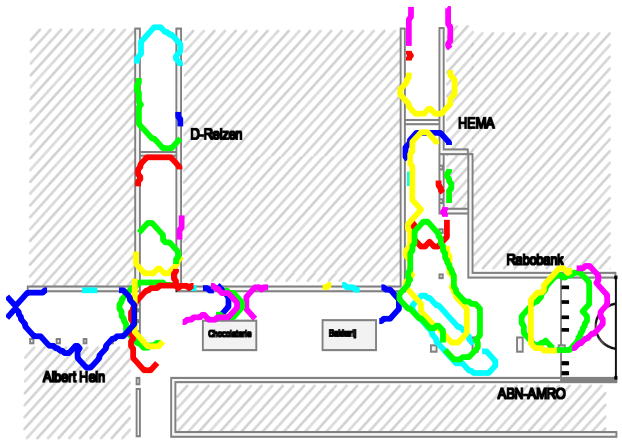


Figure 8. All automatically computed camera regions overlaid on the map of the shopping mall. (The figure is best viewed in colour.)

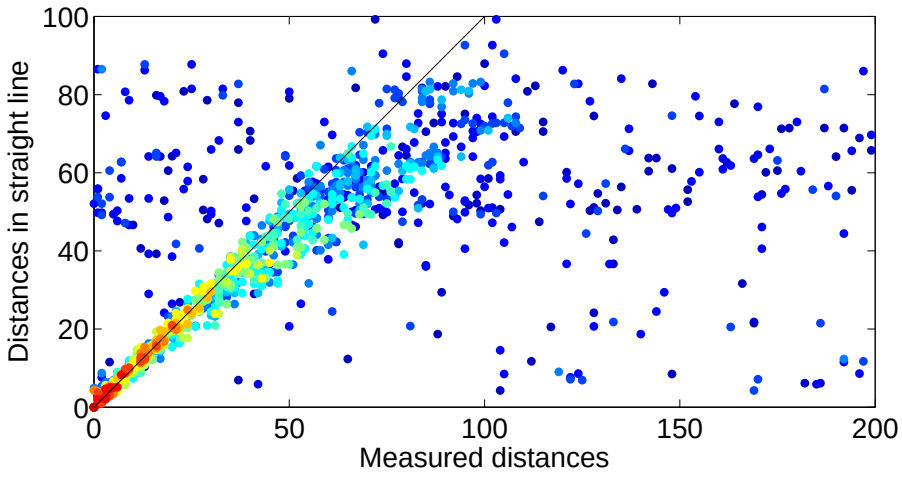


Figure 9. The maximum correlation values between all pairwise regions plotted against the measured distance and the ground truth straight line distance. Red indicates high correlation values. (The figure is best viewed in colour.)

correctly. However, the part at the lower right with camera regions 7 and 12 has essentially an undetermined angle, because there is only a single route between these two regions and other regions in the network. Also note that the orientation of the whole graph can not be determined, and we have rotated it manually for better visual comparison with the ground truth floorplan. The topology of the regions (i.e. positions in the area that originate from regions in the camera views) can be converted into a topology for the actual camera positions (i.e. the camera centres). This topology can then be used for coordinating observations of persons in the camera network, for instance predicting where persons will reappear after observing them in a specific camera (at arbitrary image position).

4. CONCLUSION

In this paper, we have presented methods for intra-camera calibration and inter-camera topology computation.

For the intra-camera parameters, we have introduced a new calibration method based on the pixel height of person detections in video. The average height versus distance distribution can be used for the computation of

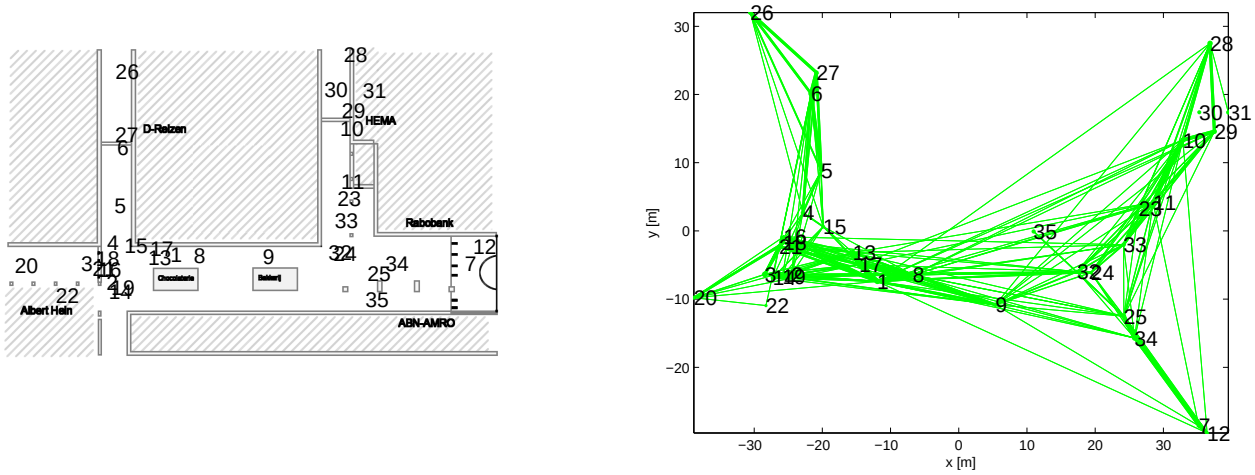


Figure 10. The actual camera region floorplan in the shopping mall (left), and the estimated camera region network with the inter-camera topology method. (The figure is best viewed in colour.)

the camera's focal length, tilt angle and height. Experiments with pedestrian detections on surveillance footage showed that the proposed method is able to give accurate parameter estimates for well fitted person detections. The estimated camera parameters result in small measurement errors for scene properties like object heights and distances. Furthermore, it was observed that for larger tilt angles and persons walking along the camera's viewing direction, the detections can become oversized and need to be corrected based on the walking pattern. Future work includes incorporating this walking pattern in the detection selection. In addition, the method could be extended with automatic lens distortion estimation, for which the height distribution across the image width at the centre can be used.

The inter-camera topology has been derived with a new method based on predicted passing sequences of persons. Experimental results on surveillance footage showed that the correlation between predicted and measured passing sequences is a robust measure for finding the distance between cameras. Next to the mutual camera distance, the proposed method is able to find any time offset between two non-synchronized cameras. The computed topology was shown to be graphically similar to the actual camera floorplan. Future work will focus on an improvement of the automatic entrance/exit region extraction in the camera views.

5. ACKNOWLEDGEMENTS

This work was conducted in the project 'Passive Sensors' in the Dutch top sector High Tech Systems and Materials and the EU FP7 project TACTICS. The authors acknowledge the "Diensten Centrum Beveiliging" (DCB) in Utrecht for providing the fieldlab facilities and support.

APPENDIX A. DERIVATION OF CAMERA PARAMETERS

In this appendix we derive the equations 2, 3 and 4 from Section 2.2.

Consider the camera setup as shown in Fig. 3. The origins of the world and camera coordinate frame here coincide in point **C**. The X-axis runs perpendicular to the YZ plane and is not used in our derivation (using $X = 0$ for the world points). The camera is looking downward with a tilt angle $0 < \theta < 90°$, and is h_{cam} metres above the ground plane level. A person of height h_{person} is observed by the camera. The camera coordinate frame is rotated w.r.t. the world frame, where the Y_{cam} axis runs parallel to the image plane and the Z_{cam} axis points in the direction of the principal axis. In this setup, the coordinates of a world point $\mathbf{X} = (0, Y, Z)^\top$ can be transformed to camera coordinates by

$$\begin{pmatrix} X_{\text{cam}} \\ Y_{\text{cam}} \\ Z_{\text{cam}} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \begin{pmatrix} 0 \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ Y\cos\theta - Z\sin\theta \\ Y\sin\theta + Z\cos\theta \end{pmatrix}, \tag{5}$$

and the corresponding y-pixel coordinate of this point is

$$y = \frac{fY_{\text{cam}}}{Z_{\text{cam}}} = \frac{fY\cos\theta - fZ\sin\theta}{Y\sin\theta + Z\cos\theta} \ . \tag{6}$$

We will denote the pixel height of a person in the image plane by Δy, and its value follows from the person's Y_{head} and Y_{feet} positions by

$$\begin{aligned}
\Delta y &\equiv y_{\text{feet}} - y_{\text{head}} && (7)\\
&= \frac{fY_{\text{feet}}\cos\theta - fZ\sin\theta}{Y_{\text{feet}}\sin\theta + Z\cos\theta} - \frac{fY_{\text{head}}\cos\theta - fZ\sin\theta}{Y_{\text{head}}\sin\theta + Z\cos\theta} && (8)\\
&= \frac{fh_{\text{cam}}\cos\theta - fZ\sin\theta}{h_{\text{cam}}\sin\theta + Z\cos\theta} - \frac{f(h_{\text{cam}} - h_{\text{person}})\cos\theta - fZ\sin\theta}{(h_{\text{cam}} - h_{\text{person}})\sin\theta + Z\cos\theta} && (9)\\
&= \frac{fZh_{\text{person}}}{(h_{\text{cam}}\sin\theta + Z\cos\theta)\left((h_{\text{cam}} - h_{\text{person}})\sin\theta + Z\cos\theta\right)} \ . && (10)
\end{aligned}$$

It is required that we know the change of height Δy versus feet position y_{feet}. In order to find this quantity we first determine the rate of change w.r.t. Z. From straightforward manipulation of Eq. 10 it follows that

$$\frac{\partial \Delta y}{\partial Z} = \frac{fh_{\text{person}}\left(h_{\text{cam}}(h_{\text{cam}} - h_{\text{person}})\sin^2\theta - Z^2\cos^2\theta\right)}{(h_{\text{cam}}\sin\theta + Z\cos\theta)^2\left((h_{\text{cam}} - h_{\text{person}})\sin\theta + Z\cos\theta\right)^2} \tag{11}$$

and from Eq. 6 that the derivative of the feet position y_{feet} versus distance Z is

$$\begin{aligned}
\frac{\partial y_{\text{feet}}}{\partial Z} &= \frac{\partial}{\partial Z}\frac{fh_{\text{cam}}\cos\theta - fZ\sin\theta}{h_{\text{cam}}\sin\theta + Z\cos\theta} && (12)\\
&= \frac{-fh_{\text{cam}}}{(h_{\text{cam}}\sin\theta + Z\cos\theta)^2} \ , && (13)
\end{aligned}$$

so that the sought derivative of person height versus feet position is

$$\begin{aligned}
\frac{\partial \Delta y}{\partial y_{\text{feet}}} &= \frac{\partial \Delta y}{\partial Z}\frac{\partial Z}{\partial y_{\text{feet}}} && (14)\\
&= \frac{-h_{\text{person}}\left(h_{\text{cam}}(h_{\text{cam}} - h_{\text{person}})\sin^2\theta - Z^2\cos^2\theta\right)}{h_{\text{cam}}\left((h_{\text{cam}} - h_{\text{person}})\sin\theta + Z\cos\theta\right)^2} \ . && (15)
\end{aligned}$$

In order to use the derived quantities above, we must analyse their values for a person at a known distance Z. Only at the principal point there is a relation of Z with the camera parameters, namely $Z = h_{\text{cam}}/\tan\theta = h_{\text{cam}}\cos\theta/\sin\theta$. We therefore evaluate Δy and $\partial\Delta y/\partial y_{\text{feet}}$ at the principal point (where it holds that $y_{\text{feet}} = 0$) and get

$$\begin{aligned}
\Delta y_{\text{pp}} &\equiv \Delta y|_{principal\ point} = -y_{\text{head}}|_{principal\ point} && (16)\\
&= \frac{-f(h_{\text{cam}} - h_{\text{person}})\cos\theta + f\frac{h_{\text{cam}}\cos\theta}{\sin\theta}\sin\theta}{(h_{\text{cam}} - h_{\text{person}})\sin\theta + \frac{h_{\text{cam}}\cos\theta}{\sin\theta}\cos\theta} && (17)\\
&= \frac{-f(h_{\text{cam}} - h_{\text{person}})\cos\theta\sin\theta + fh_{\text{cam}}\cos\theta\sin\theta}{h_{\text{cam}}\sin^2\theta - h_{\text{person}}\sin^2\theta + h_{\text{cam}}\cos^2\theta} && (18)\\
&= \frac{fh_{\text{person}}\cos\theta\sin\theta}{h_{\text{cam}} - h_{\text{person}}\sin^2\theta} \ , && (19)
\end{aligned}$$

and for the height derivative at the principal point

$$\frac{\partial \Delta y}{\partial y_{\text{feet}}}_{\text{pp}} \equiv \frac{\partial \Delta y}{\partial y_{\text{feet}}}|_{principal\ point} \tag{20}$$

$$= \frac{-h_{\text{person}}\left(h_{\text{cam}}(h_{\text{cam}}-h_{\text{person}})\sin^2\theta - h_{\text{cam}}^2\frac{\cos^4\theta}{\sin^2\theta}\right)}{h_{\text{cam}}\left((h_{\text{cam}}-h_{\text{person}})\sin\theta + h_{\text{cam}}\frac{\cos^2\theta}{\sin\theta}\right)^2} \tag{21}$$

$$= \frac{-h_{\text{person}}\left((h_{\text{cam}}-h_{\text{person}})\sin^4\theta - h_{\text{cam}}\cos^4\theta\right)}{\left((h_{\text{cam}}-h_{\text{person}})\sin^2\theta + h_{\text{cam}}\cos^2\theta\right)^2} \tag{22}$$

$$= \frac{-h_{\text{person}}\left(-h_{\text{person}}\sin^4\theta + h_{\text{cam}}\sin^2\theta - h_{\text{cam}}\cos^2\theta\right)}{\left(h_{\text{cam}}\sin^2\theta + h_{\text{cam}}\cos^2\theta - h_{\text{person}}\sin^2\theta\right)^2} \tag{23}$$

$$= \frac{-h_{\text{person}}\left(\left(h_{\text{cam}}-h_{\text{person}}\sin^2\theta\right)\sin^2\theta - h_{\text{cam}}\cos^2\theta\right)}{\left(h_{\text{cam}}-h_{\text{person}}\sin^2\theta\right)^2} \tag{24}$$

where in the fore last equality we have used the identity $\sin^4\theta - \cos^4\theta = \sin^2\theta - \cos^2\theta$.

Now we substitute the camera height h_{cam} from Eq. 19 into Eq. 24 which yields

$$\frac{\partial\Delta y}{\partial y_{\text{feet}}}_{\text{pp}} = \frac{-h_{\text{person}}\Delta y_{\text{pp}}^2\left(\frac{f h_{\text{person}}\cos\theta\sin\theta}{\Delta y_{\text{pp}}}\sin^2\theta - \left(\frac{f h_{\text{person}}\cos\theta\sin\theta}{\Delta y_{\text{pp}}} + h_{\text{person}}\sin^2\theta\right)\cos^2\theta\right)}{f^2 h_{\text{person}}^2\cos^2\theta\sin^2\theta}\ , \tag{25}$$

which can be written as

$$f^2 h_{\text{person}}^2\cos^2\theta\sin^2\theta\frac{\partial\Delta y}{\partial y_{\text{feet}}}_{\text{pp}} = -f h_{\text{person}}^2\Delta y_{\text{pp}}\cos\theta\sin^3\theta \tag{26}$$

$$+f h_{\text{person}}^2\Delta y_{\text{pp}}\cos^3\theta\sin\theta \tag{27}$$

$$+h_{\text{person}}^2\Delta y_{\text{pp}}^2\cos^2\theta\sin^2\theta\ . \tag{28}$$

Subsequently, we substitute the following expressions for the sine and cosine (see Fig. 3) into Eq. 28:

$$\sin\theta = \frac{-y_{\text{horizon}}}{\sqrt{f^2+y_{\text{horizon}}^2}}\ , \qquad \cos\theta = \frac{f}{\sqrt{f^2+y_{\text{horizon}}^2}}\ . \tag{29}$$

Since all terms have sines and cosines to the (combined) fourth power, the square root denominator can be factored out. After further removal of common factors f^2, h_{person}^2 and y_{horizon} this results in

$$f^2 y_{\text{horizon}}\frac{\partial\Delta y}{\partial y_{\text{feet}}}_{\text{pp}} = \Delta y_{\text{pp}}y_{\text{horizon}}^2 - f^2\Delta y_{\text{pp}} + y_{\text{horizon}}\Delta y_{\text{pp}}^2\ , \tag{30}$$

which can be solved for the focal length as shown in Eq. 2. The relation between the horizon and focal length (see Fig. 3) is then used for deriving the tilt angle as in Eq. 3. The camera height can be computed from Eq. 19 with additional information about the person's height, resulting in Eq. 4.

REFERENCES

[1] Bouma, H., Baan, J., Landsmeer, S., et al., "Real-time tracking and fast retrieval of person in multiple surveillance cameras of a shopping mall," *Proc. SPIE 8756*, (2013).

[2] Burghouts, G., Schutte, K., Hove, R., et al., "Instantaneous threat detection based on a semantic representation of activities, zones and trajectories," *Signal Image and Video Processing* 8(1), 191–200 (2014).

[3] Bouma, H., Baan, J., Burghouts, G., et al., "Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall," *Proc. SPIE 9253*, (2014).

[4] Wang, X., "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters* 34(1), 3–19 (2013).

[5] Song, M., Tao, D., and Maybank, S., "Sparse camera network for visual surveillance – A comprehensive survey," *arXiv 1302.0446,* (2013).

[6] Lv, F., Zhao, T., and Nevatia, R., "Camera calibration from video of a walking human," *IEEE Trans. on Pattern Analysis and Machine Intelligence* 28, 1513–1518 (Sept 2006).
[7] Krahnstoever, N. and Mendonca, P., "Bayesian autocalibration for surveillance," *Proc. IEEE Int. Conf. on Computer Vision*, 1858–1865 (Oct 2005).
[8] Krahnstoever, N. and Mendonca, P., "Autocalibration from tracks of walking people," *Proc. British Machine Vision Conference*, (2006).
[9] Junejo, I. and Foroosh, H., "Robust auto-calibration from pedestrians," *Int. Conf. on Advanced Video and Signal-based Surveillance*, 92–92 (Nov 2006).
[10] Liu, J., Collins, R., and Liu, Y., "Automatic surveillance camera calibration without pedestrian tracking," *Proc. British Machine Vision Conference*, 117.1–117.11 (2011).
[11] Hartley, R. and Zisserman, A., [*Multiple View Geometry in Computer Vision*], Cambridge University Press (2003).
[12] Mohedano, R. and Garcia, N., "Capabilities and limitations of mono-camera pedestrian-based autocalibration," *Proc. IEEE Int. Conf. on Image Processing*, 4705–4708 (Sept 2010).
[13] Richardson, E., Peleg, S., and Werman, M., "Scene geometry from moving objects," *Int. Conf. on Advanced Video and Signal-based Surveillance*, 13–18 (Aug 2014).
[14] Dollar, P., Belongie, S., and Perona, P., "The fastest pedestrian detector in the west," *Proc. British Machine Vision Conference*, (2010).
[15] Makris, D., Ellis, T., and Black, J., "Bridging the gaps between cameras," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, II–205–II–210 (June 2004).
[16] Tieu, T., Dalley, G., and Grimson, W., "Inference of non-overlapping camera network topology by measuring statistical dependence," *Proc. IEEE Int. Conf. on Computer Vision*, 1842–1849 (Oct 2005).
[17] Zou, X., Bhanu, B., Song, B., and Roy-Chowdhury, A., "Determining topology in a distributed camera network," *Proc. IEEE Int. Conf. on Image Processing*, V – 133–V – 136 (Sept 2007).
[18] Pflugfelder, R. and Bischof, H., "People tracking across two distant self-calibrated cameras," *Int. Conf. on Advanced Video and Signal-based Surveillance*, (2007).
[19] Gilbert, A. and Bowden, R., "Incremental, scalable tracking of bojects inter camera," *Computer Vision and Image Understanding* 111(1), 43–58 (2008).
[20] Loy, C., Xiang, T., and Gong, S., "Multi-camera activity correlation analysis," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1988–1995 (2009).
[21] Kuo, C. and Huang, C., "Inter-camera association of multi-target tracks by on-line learned appearance affinity models," *Proc. European Conf. on Computer Vision*, 383–396 (2010).
[22] Wang, X., Tieu, K., and Grimson, W., "Correspondence-free activity analysis and scene modeling in multiple camera views," *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32, 56–71 (Jan 2010).
[23] Nam, Y., Rho, S., and Park, J., "Inference topology of distributed camera networks with multiple cameras," *Multimedia Tools and Appl.* 67(1), 289–309 (2013).
[24] Sun, X., Chang, F., and Dong, W., "Particle filter-based object tracking and handover in disjoint view multi-cameras," *Foundations and Practical Appl. Cogn. Syst. Inf. Proc. 215*, 57–68 (2013).
[25] Bedagkar-Gala, A. and Shah, S., "A survey of approaches and trends in person re-identification," *Image and Vision Computing* 32(4), 270–286 (2014).
[26] Chen, X., Huang, K., and Tan, T., "Object tracking across non-overlapping views by learning inter-camera transfer models," *Pattern Recognition* 47(3), 1126–1137 (2014).
[27] Dick, A., Hengel, A., and Detmold, H., "Large-scale camera topology mapping: application to re-identification," *Person re-identification*, 391–411 (2014).
[28] Lee, K., Chu, C., Lee, Y., et al., "Consistent human tracking over self-organized and scalable multiple-camera networks," *Distributed Embedded Smart Cameras*, 189–209 (2014).
[29] Yin, F., Velastin, S., Ellis, T., and Makris, D., "Learning multi-planar scene models in multi-camera videos," *IET Computer Vision* 9(1), 25–40 (2014).
[30] Zhang, H., Cui, J., Wang, P., and Zheng, S., "Activity-based scene decomposition for topology inference of video surveillance network," *J. Elec. Computer Eng.,* (2014).
[31] Kruskal, J., "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika* 29(1), 1–27 (1964).