Smart Phone Accuracy of Multi-Camera Pedestrian Tracking in Overlapping Fields of View

STEFFEN BUSCH[1]

Abstract: In this paper, we analyze the suitability of event mapping via smart phones by evaluating the accuracy of pedestrian tracking. We employ a multiple view tracking and bundle adjustment and recover the scale by an alignment to reference points measured via a total station. The use of multiple perspectives by loosely time synchronized smart phones, enables a least squares assignment of detections while improving the pose estimation of pedestrians and reducing occlusions. Finally, we used a total station to track a pedestrian to evaluate our results and show that the photogrammetric smart phone tracking accuracy of 20 cm is suitable for event mapping in the daily traffic situations.

1 INTRODUCTION

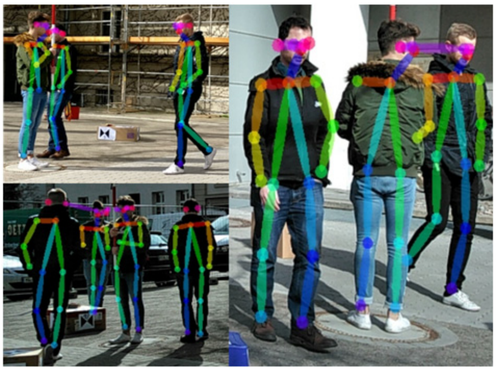


Fig. 1: Three time synchronized images of a scene, taken from three different camera viewpoints. Estimated person pose is overlaid. The bottom left and right images show a mismatch in the head association to illustrate that this detection, too, fails sometimes.

Nowadays, with the explosive development of technology for the internet of things, various sensors record lots of information in our world. This work is about the pedestrian tracking accuracy examination by exploiting wide spread smart phone sensors. By use of smart phones we will be able to comprehensively map the movement of pedestrians with multiple mobile devices, we call agents. As we demonstrated in our previous work (BUSCH et al. 2016), the movements of objects can be utilized to generate dynamic maps. These maps provide information about the real paths (trajectories) of pedestrians, when and where events occur and, similar to floating car data, traffic information.

The resulting data can be used to generate information about periodic events and especially to detect anomalies such as traffic accidents or crowded scenes. Such dynamic maps can improve the navigation in cities and be a useful tool for city planning (FAYAZI et al. 2015). The more detailed mapping of the spatiotemporal traffic behavior will play a significant role in future mobility solutions. Thus, pedestrian trajectories are a valuable source for such new kinds of maps. The tracking information could be obtained with high frequency by the daily traffic of the future, when more traffic participants are equipped with appropriate sensors. As of now, a comprehensive data collection is possible by traffic control or smart phone cameras, for instance. This is achieved by following the concept of holding maps

[1] Leibniz Universität Hannover, Institute of Cartography and Geoinformatics, Appelstraße 9a, D-30167 Hannover, E-Mail: Steffen.Busch@ikg.uni-hannover.de

up-to-date by using crowd-sensing information of the daily traffic. For that reason we use smart phones as ”all-in-one” sensors and analyze their accuracy. In order to investigate the pedestrian tracking from multiple viewpoints, we used a multi-camera network consisting of smart phone cameras, to analyze the tracking accuracy. This is relevant in maps due to the necessity for reliable 3D information like trajectory planning and obstacle handling. The real world tracking results rely on the position and orientation of the cameras, which we determined by a point cloud alignment, since this method can locate independent cameras and enables a crowd sensed tracking. During the last decade, we have witnessed an explosion in Computer Vision techniques for pedestrian detection and tracking (LEAL-TAIXÉ et al. 2016; LINDER et al. 2016; ZHANG et al. 2016). This task is critical for surveillance, autonomous driving and robotics applications. Even though pedestrian detection and tracking have attracted a lot of researchers, it still poses many challenges due to the alteration of pedestrians' pose, scale, or changes in the scene illumination. Additionally, in a crowded scene, pedestrians are usually occluded by others, which is a problem that cannot be solved effectively by single camera views. Combining images from multiple viewpoints can provide a more comprehensive knowledge of a scene, in which missing information in a particular view can be supplemented by the others (MITTAL & DAVIS 2003). This method not only helps to overcome the limitations of monocular camera views, but also improves the robustness of the 3D position estimation during the pedestrian tracking procedure. However, it is difficult to fuse information from different camera views because of uncertainties in detection and pedestrian correspondence ambiguities. To tackle this problem, we propose a framework to track pedestrians in 3D on a ground plane using images from multiple cameras. Our method extends the recursive dynamic Bayes network multi-person tracker, which was introduced by KLINGER et al. (2015) to a scenario of three cameras. Besides that, we propose a method for finding the correct combination of detections from different views. For this purpose, we combine detections of different perspectives by spatial intersection and a Gauss-Markov method. Then we calculate the probability of the resulting combinations and filter outliers. Afterwards, we select the most likely combinations and utilize the prediction information of the Kalman Filter for the association process. Thus, we strive to enhance the reliability and precision of trajectories. We test our approach on a database produced by ourselves, in which areas of interest were captured by three smart phone cameras placed at three different viewpoints. One pedestrian tracking trajectory is compared against the ground truth trajectory determined using a total station (tachymeter). Figure 2 illustrates the alignment problem of multiple detections from different perspectives. The derivation from the detectors leads to several combination possibilities by spatial intersection between the viewpoints. Our approach is looking for the correct combination of detections by filtering the set of all possible combinations.

The rest of the paper is structured as follows: in Section 2 we discuss previous relevant research. Section 3 describes our method in detail. Section 4 illustrates our experiment with the captured scene. After an evaluation of our approach in Section 5 we sum up with conclusions in Section 6.

2 Related Work

This section starts with an overview of (event) mapping approaches, since the comprehensive mapping of events has inspired our work. Afterwards, different detection approaches are presented due to their high influence on the tracking approaches. Finally, various tracking approaches are discussed.

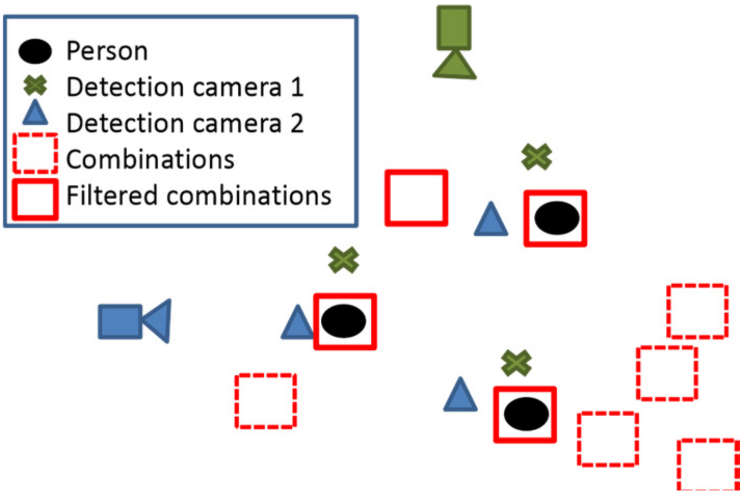


Fig. 2: Filtering of all possible combinations generated by the least square assignment of spatial intersection of different angles.

Today floating car data (FCD) is used for more effective path planning, avoiding congestions and optimizing transportation systems. Various works (RAZA & ZHONG 2017; TREIBER & KESTING 2013) predict traffic flow information based on trajectory analysis to make traffic faster, safer and more environmentally friendly. These approaches analyze trajectories generated by camera observations or by ordinary Global Navigation Satellite System (GNSS) data matched to a map. In addition to FCD, even maps and much more information are generated and updated by crowd-sensing data (HAKLAY & WEBER, 2008; HU et al., 2017). LANDSIEDEL & WOLLHERR (2017) calculated a root mean square error of around 1.5m between open street map, which is generated by using ordinary GNSS receivers, and a laser scanner reference map. GNSS trajectories generated by crowd sourcing are a widespread source to map much more information. For example ROETH et al. (2017), RUHHAMMER et al. (2017) and DURAN et al. (2016) automatically mapped road network graphs by trajectory analysis. ZOURLIDOU & SESTER (2015) and EFENTAKIS et al. (2017) used GPS tracks to enrich network graphs with information about turning behavior. Furthermore RUHHAMMER et al., (2017) and FAYAZI et al. (2015) generate higher detailed information about the spatial-temporal behavior at the network graph through traffic light sequences (TLS). In addition, they identify TLS as valuable information for traffic flow optimization. Furthermore, WANG et al. (2016) show that with the knowledge of TLS the travel time could be reduced by about 36%. All of the trajectory analysis approaches depend on trajectory precision. ATIA et al. (2017) showed that at least an accuracy of around 0.90 m is required for match detection to lane accurate maps. The precision of “GNSS only solutions” are not suitable for crowd sourcing because the low cost sensors’ error varies between 5 m and 22 m (DURAN et al., 2016), but high cost solutions are expensive for a comprehensive and high frequency use. (KNOOP et al., 2017) used GPS Precise Point Positioning to adjust low cost measurements and reached a precision of 1.2 m with a reliability of 95%. Another more accurate low cost approach is presented by VIVACQUA et al. (2017), their vision based self-localization reached a precision of 0.06 m average and 0.54 m maximal error. We use these quality measures to evaluate smart phones as a valuable crowd sensing tracking sensor. Next, we discuss the previous studies concerning pedestrian detection and tracking.

Pedestrian detection and tracking have been studied intensively during the last 20 years and their performances have been boosted continuously. Currently, there are three principal families of methods for detecting and localizing pedestrians in images. The first group of algorithms extracts hand-crafted features such as histogram oriented gradient (HOG), integral channel feature (ICF) or deformable part model (DMP) (DOLLÁR et al., 2009) of relevant objects, followed by a supervised classifier like SVM (DALAL & TRIGGS, 2005; VIOLA & JONES, 2001; FREUND & SCHAPIRE, 1997). The other method trend utilizes the Convolutional Neural Network (CNN) which is able to automatically obtain the optimum and high level features from raw pixels by training a deep network with great amounts of positive and negative training samples (CAO et al., 2016; CORDTS et al., 2016; DENG et al., 2009; LIN et al., 2014; REDMON & FARHADI, 2016). These two classes of methods deliver comparable results on different benchmark databases, such as KITTI Vision Benchmark Suite (GEIGER et al., 2012), ETH (ESS et al., 2008), MOTChallenge (LEAL-TAIXÉ et al., 2015). Another class of detector approaches for static cameras uses background modelling (ST-CHARLES et al., 2015) to detect a person by detecting differences between dynamic and static background layers and a current frame.

Using detections in each image frame, the tracking can be accomplished by associating pedestrians between two consecutive image frames. Most of the state-of-the-art methods employ the tracking-by-detection pipeline to track pedestrians, both in 2D and 3D spaces, and achieved impressive results (CHOI et al., 2013; KLINGER et al., 2017; MILAN et al., 2013). KLINGER et al., (2017) utilized the HOG-SVM to detect pedestrians in an image sequence. Next, the pedestrian association was performed using linear programming. They employed the Kalman filter and a dynamic Bayesian network to predict and correct the pedestrian trajectories as well. Moreover, they did not consider each pedestrian movement independently, but took into account interactions and affectations of other movements in the scene as well.

Our tracking method for multiple viewpoints is an extension of KLINGER'S (2017) approach. However, we take it a step further by replacing the HOG-SVM with deep learning detectors to obtain improved detection accuracy. We also include a multi view alignment module to determine the corresponding pedestrian across all viewpoints. Despite many techniques have been developed to solve the pedestrian tracking challenges, tracking from a single camera viewpoint usually fails in case of occlusion and is less accurate due to lacking information concerning the whole scene. Several approaches have been studied to leverage multi-view object tracking. To track pedestrians in crowded scenes, KHAN & SHAH (2006) performed a planar homography constraint to clarify the occlusion and localize positions of pedestrians based on the measured foot blobs. Both BERCLAZ et al. (2011) and LEAL-TAIXE et al. (2012) treated the multi-view tracking problem as a flow optimization task. Utilizing the relationship of knowledge obtained from multiple view point images by joining spare learning, HONG et al. (2013) could extract more sufficient information to boost the performance of his particle-filter-based tracker. XU et al. (2016) adopted a hierarchical composition model to generate the object trajectories through maximizing a posterior. While the existing approaches achieved remarkable improvements employing different approaches, they are more or less imposed on appearance resemblance, movement smoothness, sparsity, 3D position coincidence, etc., of pedestrians over temporal and spatial domains. In the same manner, we take those cues into the tracker. Our tracking approach differs from high precision multi-camera tracking approaches such as used for motion capturing systems for film studios since we do not apply our approach in a closed room with defined background for Chroma keying with high precision time synchronized cameras, a full bundle adjustment over all cameras with unlimited observations and controlled light conditions.

We focus on the multi view alignment task to reduce the effect of occlusion, in which the corresponding pedestrians are determined with assumption of 3D position coincident over multiple viewpoints. Moreover, we perform our tracking in 3D world space and compare our results to a ground truth trajectory to get a quantitative measurement of accuracy we can achieve in realistic applications.

3 Multi-View Tracking

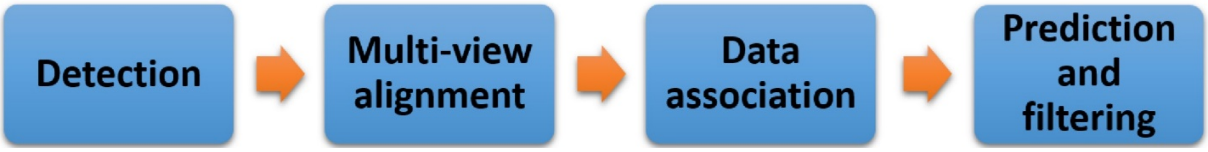


Fig. 3: The framework of our proposed tracker

We extend the conventional tracking approach from KLINGER et al. (2017) by use of different perspectives from a multi-camera network. The different viewpoints enable a more precise 3D position estimation in comparison to a single mono camera. Moreover, they make the tracking much more robust against occlusion. The framework of our approach includes four primary phases (see Figure 3): first, the detection, in which pedestrians are recognized and localized separately in each image. Secondly, the alignment task is executed to find the corresponding pedestrians between different views. Next, the detections are concatenated across frames in the data association phase. Finally, the trajectories of pedestrians are smoothed by applying a Kalman filter for the position of detections and predicted locations inferred from previous image frames. This section will first explain the localization methods since the tracking is based on the known camera positions. Afterwards, each step is explained in more detail in the corresponding chapter.

3.1 Pose Estimation via Point Cloud Alignment

To determine the camera positions, we use a point cloud alignment because this method allows for an independent localization for each camera. The poses only depend on a point cloud map and a short image sequence of a moving camera. For the alignment we calculate a point cloud via bundle adjustment and align it to the reference points from the total station. For this we pick corresponding points manually for a Four Point Congruent Sets (4PCS) (AIGER et al. 2008) registration algorithm. Thus, a transformation including the scale between the coordinate frame of the bundle adjustment and the coordinate frame of the total station is calculated. With this pose information we identify the ground plane. This ground plane is used to project every detection onto the ground and approximate the real world position for each detection.

3.2 Pedestrian Detection

For the detection, we adapt two different deep learning based methods which were recently published and can efficiently perform the pedestrian detection task. On one side, YOLO: Real-Time Object Detection (REDMON & FARHADI 2016) produces bounding boxes of pedestrians; we interpret the center of the bottom bounding box line as a foot point for the tracking. On the other side, the 2D pose estimation approach (2DPE) (CAO et al. 2016; LIN et al. 2014) delivers the skeletons of detected pedestrians, whereby we combine the neck x-position and lowest ankle y-position in order to precisely detect the foot point of pedestrians. Particularly, only the pre-trained networks provided by the authors are used and no domain adaptation to our newly generated data set is done since the models generalized well. Three images from three different viewpoints with overlaid pose skeletons are depicted in Figure 1.

3.3 Multi-View Alignment

In our case, the problem of data association expands by the use of multiple cameras. To simplify the association process, we combine the detections of the same persons from different views before starting to process. Thus, we create all possible combinations from the detections of different views and assume that each combination describes the same person. We then optimize the detected 3D position of that person by correcting the image coordinates of the detections so that they describe the same point in 3D. Based on this spatial intersection by using the Gauss-Markov model (2), with the design matrix A (1), we identify the correct combinations, which have the smallest residuals. To find that smallest residual independent of camera count, we use the square improvements (3) and the chi square function (4). In more detail, we use the Jacobian matrix of the object (world) coordinates X, Y and Z of the collinearity equation (KRAUS 2004) and use the assumption that Z is constant to determine the design matrix

$$A = \begin{bmatrix} -\frac{f_0(N_0 r_{0,0,0} - \xi_0\, r_{0,2,0})}{N_0^2} & -\frac{f_0(N_0\, r_{0,0,1} - \xi_0\, r_{0,2,1})}{N_0^2} & -\frac{f_0(\, N_0\, r_{0,0,2} - \xi_0\, r_{0,2,2})}{N_0^2} \\ -\frac{f_0(\, N_0 r_{0,1,0} - \eta_0\, r_{0,2,0})}{N_0^2} & -\frac{f_0(\, N_0\, r_{0,1,1} - \eta_0\, r_{0,2,1})}{N_0^2} & -\frac{f_0(\, N_0\, r_{0,2,1} - \eta_0\, r_{0,2,2})}{N_0^2} \\ \vdots & \vdots & \vdots \\ -\frac{f_i(N_i r_{i,0,0} - \xi_i\ r_{i,2,0})}{N_i^2} & -\frac{f_i(\&N_i\, r_{i,0,1} - \xi_i\, r_{i,2,1})}{N_i^2} & -\frac{f_i(N_i\, r_{i,0,2} - \eta_i\, r_{i,2,2})}{N_i^2} \\ -\frac{f_i(\, N_i r_{i,1,0} - \eta_i\, r_{i,2,0})}{N_i^2} & -\frac{f_i(\, N_i\, r_{i,1,1} - \eta_i\, r_{i,2,1})}{N_i^2} & -\frac{f_i(N_i\, r_{i,2,1} - \eta_i\, r_{i,2,2})}{N_i^2} \\ 0 & 0 & 1 \end{bmatrix}. \qquad (1)$$

Where: f_i = focal length of camera i

$$\mathrm{N_i} = \mathrm{r}_{i,2,0}(\mathrm{X} - \mathrm{T_{i,x}}) + \mathrm{r_{i,2,1}}(\mathrm{Y} - \mathrm{T_{i,y}}) + \mathrm{r_{i,2,2}}\,(\mathrm{Z} - \mathrm{T_{i,z}})$$

$$\xi\mathrm{i} = \mathrm{r_{i,0,0}}\,(\mathrm{X} - \mathrm{T_{i\,,x}}) + \mathrm{r_{i,0,1}}(\mathrm{Y} - \mathrm{T_{i,y}}) + \mathrm{r_{i,0,2}}(\mathrm{Z} - \mathrm{T_{i,z}})$$

$$\eta\mathrm{i} = \mathrm{r_{i,1,0}}(\mathrm{X} - \mathrm{T_{(i,x)}}) + \mathrm{r_{i,1,1}}(\mathrm{Y} - \mathrm{T_{i,y}}) + \mathrm{r_{i,1,2}}(\mathrm{Z} - \mathrm{T_{i,z}})$$

$\mathrm{T_{i,x|y|z}}$ = world coordinat of camera i

r_i = orientation of camera i

We iterate the Gauss-Markov model to optimize the 3D position x' of each combination,

$$x' = (\mathrm{A^T P A})^{-1}\mathrm{A^T P l}\,. \tag{2}$$

Where:

$\mathrm{l} \quad = (\mathrm{x_0\, y_0 \ldots x_i\, y_i 0})^T$

$\mathrm{x}_i, \mathrm{y_i} \quad$ = image coordinates of detection

$$P \quad = \begin{bmatrix} \sigma_{x_0}^2 & 0 & \ldots & 0 & 0 & 0 \\ 0 & \sigma_{y_0}^2 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \ldots & \sigma_{x_i}^2 & 0 & 0 \\ 0 & 0 & \ldots & 0 & \sigma_{y_i}^2 & 0 \\ 0 & 0 & \ldots & 0 & 0 & \sigma_P \end{bmatrix}^{-1}$$

$\mathrm{\sigma_{y_i}, \sigma_{x_i}} \;=\;$ standard derivation of detector i

$\sigma_p \quad$ = standard derivation for ground Z coordinate

Thus, we improve the image observation by finding the smallest residuals v = Ax−l and use the variance propagation to find the right matches (NIEMEIER 2001). Then the sum of the squared improvements is calculated as follows:

$$\Omega = \mathrm{vTPv} \tag{3}$$

Afterwards, the chi square function is used to calculate a comparable measurement independent of the count of perspectives a. Inspired by the global test, we used the cumulative distribution of the chi square function to get a probability for the detections belonging to the same object:

$$\mathrm{P_D} = 1 - \int_0^{\Omega} \chi^2_{\mathrm{2a-2}} \tag{4}$$

Where:

a = count of different perspectives.

Finally, we filter the unlikely combinations which are smaller than a defined threshold and ensure that no detection is used twice.

3.4 Data Association

For the association from detections to trajectories we use the Mahalanobis distance (5) to weight how well a detection belongs to a track, which actual position is prediction-based on the previous walking direction:

$$\delta\mathrm{M} = \begin{pmatrix} \mathrm{x_D} & -\mathrm{x_T} \\ \mathrm{y_D} & -y_T \end{pmatrix}^{\mathrm{T}} \cdot (\Sigma_\mathrm{D} + \Sigma_\mathrm{T})^{-1} \cdot \begin{pmatrix} \mathrm{x_D} & -\mathrm{x_T} \\ \mathrm{y_D} & -\mathrm{y_T} \end{pmatrix} \tag{5}$$

Where:

$\mathrm{x_D}, \mathrm{y_D}$= coordinates of detected 2D position

$\mathrm{x_T}, \mathrm{y_T}$= coordinates of predicted 2D object position

$\Sigma_\mathrm{D}, \Sigma_\mathrm{T}$= the covariance matrices of detection and prediction

Thus, δM is used to calculate the weight $w_{D,\mathrm{T}}$ (6) for the detections belonging to a tracked pedestrian:

$$w_{D,T} \; = \; e^{-1/2\delta M} \tag{6}$$

Afterwards, we used mixed integer linear programming from BERKELAAR et al. (2004) to assign the right detections by use of a Branch-and-bound method for solving the optimization problem. Thus, we employ the constraints that every detection and trajectory could have one assignment at most.

3.5 Prediction and Filtering (Kalman Filter)

For tracking we use a dynamic Bayes Network of KLINGER et al. (2015). From each detection we calculate the two positions foot p_f and head p_h, which both have the same x-image coordinate of the neck, while y-coordinates are estimated corresponding to the coordinate of the lower ankle and the upper eye detection. The tracking is performed by an extended Kalman filter in the real world coordinate frame by tracking X, Y, Z of the foot pF and the head pH, whereby only Z differs. In addition the hidden variables for the speed [vx, vy] along X and Y are part of the tracking state 7:

$$wi,t \; = \; [X_{i\,,t}, Y_{i,t}, Z_{f,i,t}, Z_{h,i,t}, v_{x\,,i,t}, v_{y,i,t}] \tag{7}$$

Where:

i = object index
t = time index
Z_f, Z_h = Z coordinate of food and head

The measurement matrix of the Kalman filter is similar to the design matrix equation 1 but is extended by the observed head position. For this, the Z coordinates of the heads are given by the height of the detected bounding boxes.

4 Experiment

We tracked pedestrians in a network of three smartphone cameras. The cameras were placed in a triangle with an edge length of around 20 m and with a horizontal orientation. The floor plan of our setup is depicted in Figure 4.

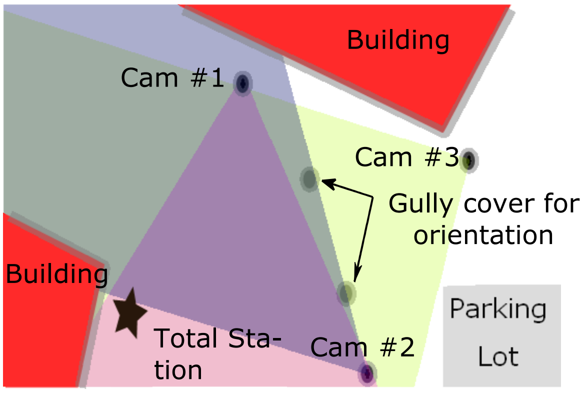


Fig. 4: Floor plan for our recorded sequences. In regions of overlapping field of views, pedestrians are visible in more than one camera.

In our experiment the agents were mounted on tripods at a height of 1.3 m above ground, a person's usual smartphone carrying height. In addition we measured a precise ground truth with 192 total station (Leicar LSM50) measurements of one pedestrian. This person carried a support rod mounted with a 360-degree prism at a height of about 3 m to prevent occlusions. For the calculation of the real world transformation, including the missing scale information, we measured 10 points with the total station and manually aligned the point cloud of the bundle adjustment. In order to focus on the accuracy of the detection and the tracking, we used only one static pose for each smartphone. The point cloud alignment could be used to localize the position of an agent based on 3D point clouds of mobile mapping systems, e.g. the 4PCS performs with a root mean square error (RMSE) of 0.027 m. Furthermore, we analyzed the influence of the transformation error in a mean pixel error of 11.3 by projecting the real world marks into the images. This relatively big mean pixel error distorted the results of less than 1 pixel standard derivation of the bundle adjustment and was added to the standard derivation of the detectors. Thus, we used a standard derivation of 40 pixels for the detected y and 30 pixels for the x image coordinate. We captured a crowded scene including 10 pedestrians for a period of two minutes. All trajectories of the tracked persons with the total station are shown in Figure 5. The pedestrians walked back and forth in the scene. We recorded many occlusions, as expertly anticipated in Figure 1. In more detail, we used a notebook to record the tracking data of the total station and time stamped the position with the processor time as soon as they arrived along the serial connection. We captured the images via action listener and time stamped this image at the arrival with the processor time of the smart phone. After we saved the image in the internal storage of the smart phone we captured the next image. Thus, the recording with full HD resolution (1920×1080 pixels) was not synchronized and the frame rate of each smart phone was not constant, but no variations were lower than 5fps. The images from different perspectives allowed for a higher tracking precision by triangulation of synchronized detections. The time synchronization was applied by the Wi-Fi internet connection of the smart phones and the notebook via the Network Time Protocol (NTP) to a global NTP sever. For the evaluation we identified a time lag of one second for the connection between total station and the notebook. Due to the irregular image capturing we used a 100 millisecond time window to synchronize the images.

5 Evaluation

The evaluation is structured in two parts, starting with the evaluation of the detection precision, including a comparison between bounding box and skeleton detectors, and the analysis of the improvements by combining multiple perspectives. Afterwards, the analysis of the tracking results is presented.

5.1 Detection Statistics

In this paragraph we discuss the detection results that we achieved with the pre-trained detectors. We evaluate every detection from our image sequence by calculating precision, recall and f-score of both YOLO and 2DPE detectors. Table 1 shows the precision, recall and f-score for the YOLO bounding box detector, while in Table 2 the results for the 2DPE are shown. The performance of the 2DPE skeleton detector on the sequences of camera 1 and camera 3 is better, while the YOLO approach delivers the higher f-score for the sequence of camera 2. Secondly, we analyze the enhancement of spatial intersection for detections. We assume that the nearest position, calculated by projecting the detection onto the ground (GP) or least square assignment of spatial intersection (OP), is the position of the reference person, called closest position assumption in the following. The next section (Tracking 5.2) will indicate that this assumption holds true in most cases. Nevertheless, the analysis will show some outliers, which appear because of an assignment of an OP of another person to the reference pedestrian. Based on the closest position assumption we compare the accuracy of the position GP and OP by their difference to the reference.

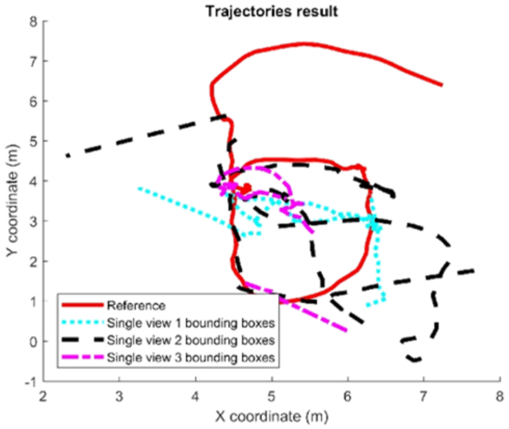


Fig. 5a: Trajectories from mono camera tracking. All trajectories belong to the pedestrian tracked by the total station

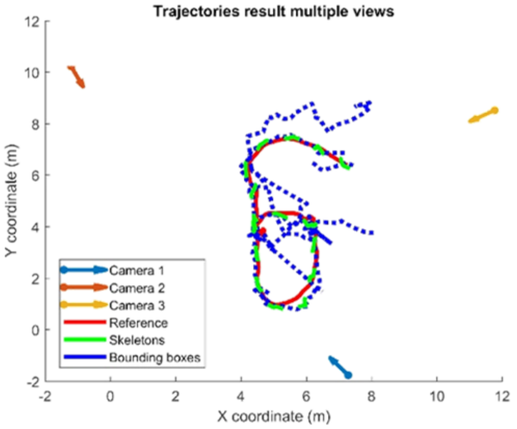


Fig. 5b: Trajectories from multi-camera tracking. All trajectories belong to the pedestrian tracked by the total station

Tab. 1: Detection results for the Yolo V2 bounding box detector

	Precision	Recall	F1-score
Camera 1	0.98	0.91	0.95
Camera 2	0.96	0.82	0.88
Camera 3	0.98	0.81	0.89

Tab. 2: Detection results for the CMU pose estimation

	Precision	Recall	F1-score
Camera 1	1	0.94	0.97
Camera 2	1	0.77	0.87
Camera 3	0.96	0.92	0.94

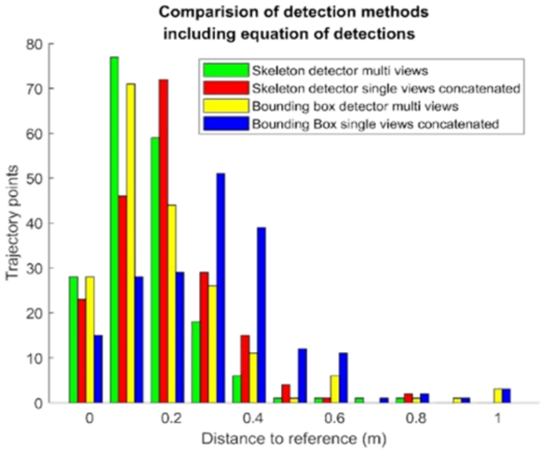
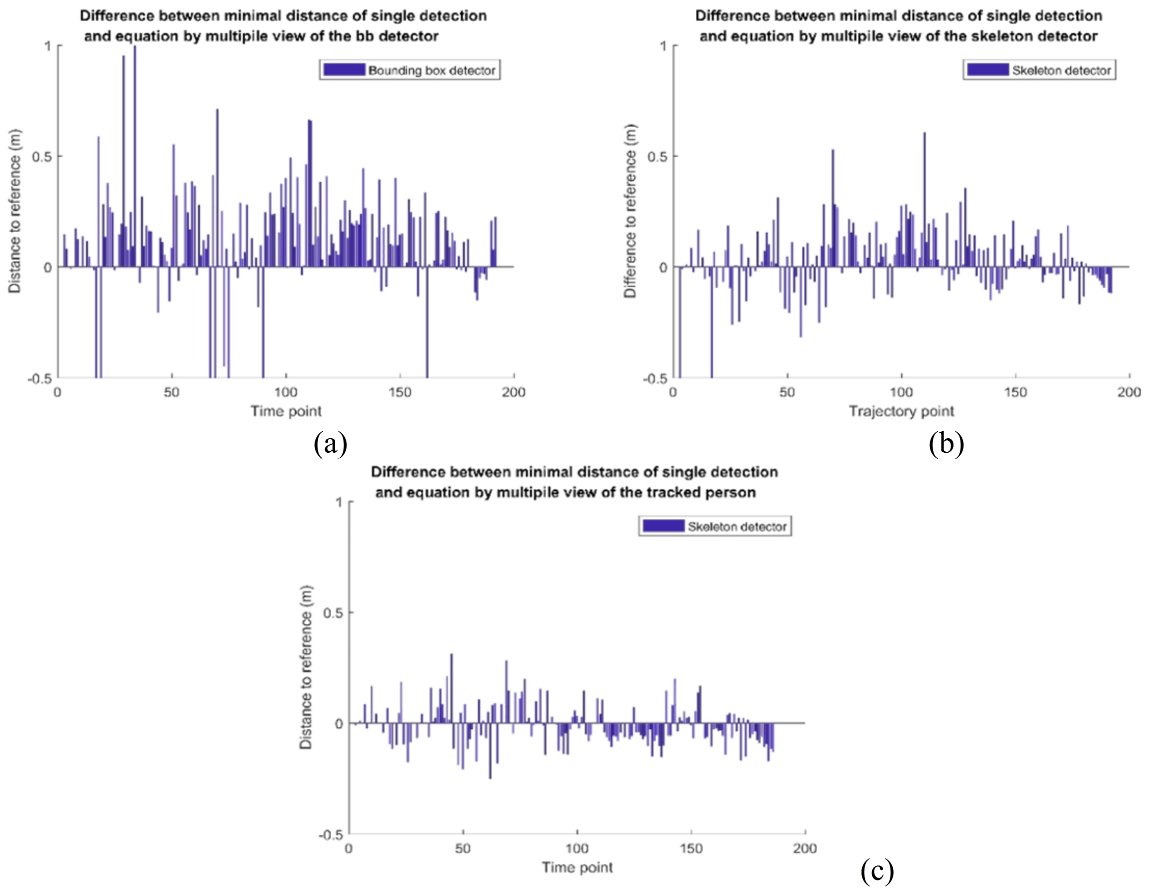


Fig. 6: Histogram of detection differences of detectors. Objective is a high number of trajectory points by a small distance to the reference

The higher accuracy of the skeleton detector is shown in Figure 6. The average precision of positions is raised from 0.31 m to 0.2 m by the spatial intersection of multiple views for the bonding box positions. This improvement is higher in contrast to the skeleton detector with an improvement of the mean accuracy from 0.2 m to 0.16 m. This difference occurs because of the higher bounding box height variation of the bounding box detector. This result shows that less accurate detections benefit more from the spatial intersection, but more variations have a fatal influence on the tracking, as will be shown in the next section (5.2 Tracking). Nevertheless, this higher accuracy shows the benefits of spatial intersection for determining positions from detections. A closer look at the positional improvement by spatial intersection is given in Figure 7 by calculating the difference between the distance to the reference of the two positions GP and OP for both detectors. A positive value means that the OP was closer to the reference and a negative value shows that the GP was closer to the reference. High negative values are explicable by comparing a spatial intersection of different persons because of the single observability of the reference pedestrian and depicted outlier. Another uncertainty of the analysis is given by the difference between the position of the camera tracked foot point and tripod point tracked by the total station. Nevertheless, Subfigure 7a shows that most of the improvements are positive with the mean of 0.101 m for the bounding box detector. Subfigure 7b shows more negative values for the skeleton detector, but the decline varies around ± 20 cm and thus is more strongly influenced by the different tracking points from cameras and the total station. Finally, Subfigure 7c shows the differences between the automatically chosen OPs and the GPs for the tracking process, which are aligned to the reference person. The fact that the outliers are missing in this case indicates that they were caused by violation of the nearest position assumption. The average change by optimization of -0.01 for the automatically chosen positions indicates enhancement potential of our filtering step. However, at least in 73 out of 186 situations an improvement at the automatically chosen positions was achieved. Besides the improvements of the estimated positions, our approach was able to summarize the detections of different views and thus performs an end-to-end tracking despite the occlusion scenarios. For the tracking, the huge set of all possible combinations was filtered to identify the right pairs of detections.

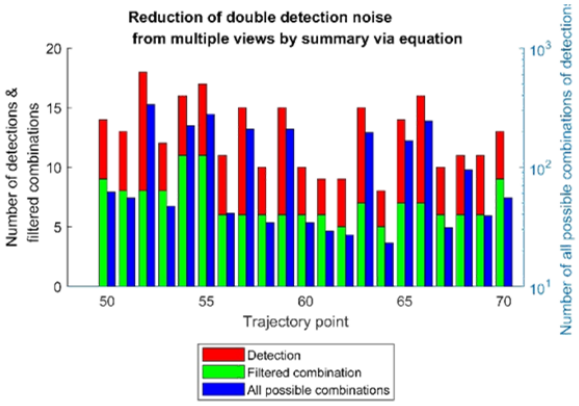


(a) (b)

(c)

Fig. 7: Differences of detection and ground truth

Figure 8 shows the fusion of the detections by summarizing the detections of different perspectives which belong to the same person. This snapshot visualizes our way from several detections through the exponential growing set of combinations to the small set of filtered positions, which reflects the real number of pedestrians at the scene.



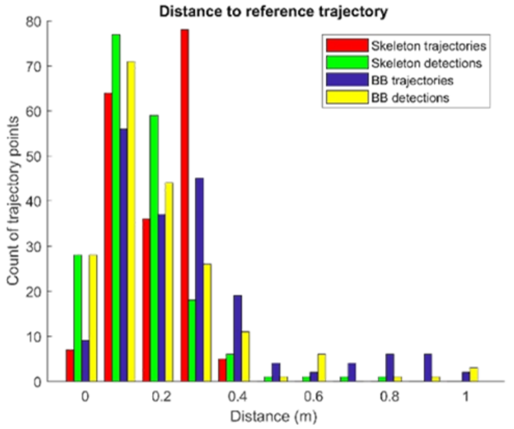


Fig. 8: Summary of detections by filtering all possible combinations of different views

Fig. 9: Histogram of the distance of tracked trajectories and nearest detected position to the reference trajectory assigned by time stamp

5.2 Tracking

After the analysis of the detection precision we analyze the tracking accuracy. The different short trajectories in Figure 5 a show that no accurate tracking is possible because of many occlusions in case of a mono view. The influence of the bounding box size variance is shown in Figure 5b, where inaccurate detections break tracks and thus new tracks start farther from the reference. That is why no end-to-end tracking was possible with the bounding box detector. For that reason, we show every trajectory by this detector, which was once assigned to the reference pedestrian in Figure 5b as dotted blue lines. In contrast to the bounding box tracking, our approach by using the skeleton detector was able to continuously track the reference pedestrian. Even in the case of occlusion situations, where the distance of pedestrians was only 30 cm - as depicted in Figure 1- we received no identity switches. Thus, an end-to-end tracking was possible and the track is shown in Figure 5b with a dashed green line. To make the end-to-end tracking result of the skeleton detector comparable to the bounding box detector with many identity switches, we used the nearest trajectory to the reference at each time step, which was assigned to the reference person once and compared their distances. Figure 9 shows the higher accuracy and the outlier resistance of tracking with use of the skeleton detector in contrast to the use of the bounding box detector. This figure shows the amount of positions over their distance to the reference. Our final result for our tracking approach via smart phones and a skeleton detector is a RMSE of 0.23 with a variance of 0.01. Besides this accuracy the small difference between the RMSE for the positions from the skeleton tracking and detector (0.23 m and 0.2 m) showed the high assignment precision of our approach. The Kalman filtering reduces the outliers (maximum position error) and shows results that are more robust. Moreover, the same RMSE analysis for the position from the bounding box detector (0.35 m and 0.31 m) demonstrates the robustness of our summary of different perspectives, even in cases of high variation at the detections. However, the variation of the bounding box detections led to more outliers and identity switches. Since skeleton detections are more precise, the identity switches are reduced from six for the bounding box detector to zero switches for the skeleton detector.

6 Conclusion

We presented a multi-view tracking approach with the overarching goal of event mapping, which was able to track a pedestrian on our dataset without identity switches despite occlusion scenarios in a crowded scene. We were able to ensure a maximal failure of 42 cm and a mean precision of 20 cm by using simple time-synchronized smartphones, a neural network detector and the spatial intersection of different viewpoints. We showed that the positions of cameras, estimated by a bundle adjustment and a point cloud alignment can be used to track pedestrians with this accuracy. In future work we will extend this approach to mobile camera networks to build a dynamic map with the long term objective of anomaly detection in dynamic maps.

7 ACKNOWLEDGEMENTS

Steffen Busch was funded by the German Science Foundation DFG within the priority program SPP1835, “Cooperative Interacting Automobiles”.

8 References

AIGER, D., MITRA, N.J. & COHEN-OR, D., 2008: 4-points Congruent Sets for Robust Surface Registration. ACM Trans. Graph., **27**(85), 1-10.

ATIA, M.M., HILAL, A.R., STELLINGS, C., HARTWELL, E., TOONSTRA, J., MINERS, W.B. & BASIR, O.A., 2017: A Low-Cost Lane-Determination System Using GNSS/IMU Fusion and HMM-Based Multistage Map Matching. IEEE Trans. Intell. Transp. Syst. **18**, 3027-3037. https://doi.org/10.1109/TITS.2017.2672541

BERCLAZ, J., FLEURET, F., TURETKEN, E. & FUA, P., 2011: Multiple object tracking using k-shortest paths optimization. IEEE Trans. Pattern Anal. Mach. Intell., **33**, 1806-1819.

BERKELAAR, M., EIKLAND, K. & NOTEBAERT, P., 2004: lpsolve: Open source (mixed-integer) linear programming system. Eindh. U Technol.

BUSCH, S., SCHINDLER, T., KLINGER, T. & BRENNER, C., 2016: Analysis of Spatio-Temporal Traffic Patterns Based on Pedestrian Trajectories. Int Arch Photogramm Remote Sens Spat. Inf Sci XLI-B2.

CAO, Z., SIMON, T., WEI, S.-E. & SHEIKH, Y., 2016: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. ArXiv Prepr. ArXiv161108050.

CHOI, W., PANTOFARU, C. & SAVARESE, S., 2013: A general framework for tracking multiple people from a moving camera. IEEE Trans. Pattern Anal. Mach. Intell., **35**, 1577-1591.

CORDTS, M., OMRAN, M., RAMOS, S., REHFELD, T., ENZWEILER, M., BENENSON, R., FRANKE, U., ROTH, S. & SCHIELE, B., 2016: The Cityscapes Dataset for Semantic Urban Scene Understanding. in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

DALAL, N. & TRIGGS, B., 2005: Histograms of Oriented Gradients for Human Detection. in: CVPR, 886-893.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. & FEI-FEI, L., 2009: ImageNet: A Large-Scale Hierarchical Image Database. in: CVPR09.

DOLLÁR, P., TU, Z., PERONA, P. & BELONGIE, S., 2009: Integral channel features.

DURAN, D., SACRISTÁN, V. & SILVEIRA, R.I., 2016: Map Construction Algorithms: An Evaluation Through Hiking Data. in: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS ’16. ACM, New York, NY, USA, 74-83. https://doi.org/10.1145/3004725.3004734

EFENTAKIS, A., GRIVAS, N., PFOSER, D. & VASSILIOU, Y., 2017: Crowdsourcing turning-restrictions from map-matched trajectories. Inf. Syst., **64**, 221-236. https://doi.org/https://doi.org/10.1016/j.is.2016.04.004

ESS, A., LEIBE, B., SCHINDLER, K., GOOL, & L. VAN, 2008: A Mobile Vision System for Robust Multi-Person Tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08). IEEE Press.

FAYAZI, S.A., VAHIDI, A., MAHLER, G. & WINCKLER, A., 2015. Traffic signal phase and timing estimation from low-frequency transit bus data. IEEE Transactions on Intelligent Transportation Systems, **16**(1), 19-28.

FREUND, Y. & SCHAPIRE, R.E., 1997: A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting.

GEIGER, A., LENZ, P. & URTASUN, R., 2012: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite, in: Conference on Computer Vision and Pattern Recognition (CVPR).

HAKLAY, M. & WEBER, P., 2008: Openstreetmap: User-generated street maps. IEEE Pervasive Comput., **7**, 12–18.

Hong, Z., Mei, X., Prokhorov, D. & Tao, D., 2013. Tracking via robust multi-task multiview joint sparse representation, in: Proceedings of the IEEE International Conference on Computer Vision, 649-656.

HU, M., ZHONG, Z., NIU, Y. & NI, M., 2017: Duration-Variable Participant Recruitment for Urban Crowdsourcing With Indeterministic Trajectories. IEEE Trans. Veh. Technol., **66**, 10271-10282. https://doi.org/10.1109/TVT.2017.2718043

KHAN, S. & SHAH, M., 2006: A multiview approach to tracking people in crowded scenes using a planar homography constraint. Comput. Vision–ECCV 2006 133-146.

KLINGER, T., ROTTENSTEINER, F. & HEIPKE, C., 2017: Probabilistic multi-person localisation and tracking in image sequences. ISPRS J. Photogramm. Remote Sens.

KLINGER, T., ROTTENSTEINER, F. & HEIPKE, C., 2015: PROBABILISTIC MULTI-PERSON TRACKING USING DYNAMIC BAYES NETWORKS. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. II-3/W5, 435-442. https://doi.org/10.5194/isprsannals-II-3-W5-435-2015

KNOOP, V.L., DE BAKKER, P.F., TIBERIUS, C.C. & VAN AREM, B., 2017: Lane Determination With GPS Precise Point Positioning. IEEE Transactions on Intelligent Transportation Systems, **18**(9), 2503-2513.

KRAUS, K., 2004: Photogrammetrie, 3rd ed, 7. Walter de Gruyter, 10785 Berlin.

LANDSIEDEL, C. & WOLLHERR, D., 2017: Road geometry estimation for urban semantic maps using open data. Adv. Robot., **31**, 282-290. https://doi.org/10.1080/01691864.2016.1250675

LEAL-TAIXÉ, L., CANTON-FERRER, C. & SCHINDLER, K., 2016: Learning by tracking: Siamese CNN for robust target association, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 33-40.

LEAL-TAIXÉ, L., MILAN, A., REID, I., ROTH, S. & SCHINDLER, K., 2015: MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. ArXiv150401942 Cs.

LIN, T.-Y., MAIRE, M., BELONGIE, S.J., BOURDEV, L.D., GIRSHICK, R.B., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P.& ZITNICK, C.L., 2014: Microsoft COCO: Common Objects in Context. CoRR abs/1405.0312.

LINDER, T., BREUERS, S., LEIBE, B. & ARRAS, K.O., 2016: On multi-modal people tracking from mobile platforms in very crowded and dynamic environments. in: IEEE International Conference on Robotics and Automation (ICRA), 5512-5519.

MILAN, A., SCHINDLER & K., ROTH, S., 2013: Detection-and trajectory-level exclusion in multiple object tracking. in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3682-3689.

MITTAL, A. & DAVIS, L.S., 2003: M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. Int. J. Comput. Vis., **51**, 189-203.

NIEMEIER, W., 2001: Ausgleichsrechnung, 3rd ed, 1. Walter de Gruyter, 10785 berlin.

RAZA, A. & ZHONG, M., 2017: Hybrid lane-based short-term urban traffic speed forecasting: A genetic approach. in: 4th International Conference on Transportation Information and Safety (ICTIS), 271-279. https://doi.org/10.1109/ICTIS.2017.8047776

REDMON, J. & FARHADI, A., 2016: YOLO9000: Better, Faster, Stronger. ArXiv Prepr. ArXiv161208242.

ROETH, O., ZAUM, D. & BRENNER, C., 2017: Extracting Lane Geometry and Topology Information from Vehicle Fleet Trajectories in Complex Urban Scenarios Using a Reversible Jump Mcmc Method. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **4**, 51.

RUHHAMMER, C., BAUMANN, M., PROTSCHKY, V., KLOEDEN, H., KLANNER, F. & STILLER, C., 2017: Automated Intersection Mapping From Crowd Trajectory Data. IEEE Trans. Intell. Transp. Syst., **18**, 666-677.

ST-CHARLES, P.-L., BILODEAU, G.-A. & BERGEVIN, R., 2015: Subsense: A universal change detection method with local adaptive sensitivity. IEEE Trans. Image Process. **24**, 359-373.

TREIBER, M. & KESTING, A., 2013: Trajectory and Floating-Car Data, in: Traffic Flow Dynamics: Data, Models and Simulation. Springer Berlin Heidelberg, Berlin, Heidelberg, 7-12. https://doi.org/10.1007/978-3-642-32460-4_2

VIOLA, P. & JONES, M., 2001: Robust Real-time Object Detection, in: International Journal of Computer Vision.

VIVACQUA, R.P.D., BERTOZZI, M., CERRI, P., MARTINS, F.N. & VASSALLO, R.F., 2017: Self-Localization Based on Visual Lane Marking Maps: An Accurate Low-Cost Approach for Autonomous Driving. IEEE Transactions on Intelligent Transportation Systems, 1-16. https://doi.org/10.1109/TITS.2017.2752461

WANG, S., DJAHEL, S., ZHANG, Z. & MCMANIS, J., 2016: Next Road Rerouting: A Multiagent System for Mitigating Unexpected Urban Traffic Congestion. IEEE Trans. Intell. Transp. Syst., **17**, 2888-2899. https://doi.org/10.1109/TITS.2016.2531425

XU, Y., LIU, X., LIU, Y. & ZHU, S.-C., 2016: Multi-view people tracking via hierarchical trajectory composition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4256-4265.

ZHANG, S., BENENSON, R., OMRAN, M., HOSANG, J. & SCHIELE, B., 2016: How far are we from solving pedestrian detection? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1259-1267.

ZOURLIDOU, S. & SESTER, M., 2015. Road regulation sensing with in-vehicle sensors. in: AGILE PhD School.