2012 Intelligent Vehicles Symposium
Alcalá de Henares, Spain, June 3-7, 2012

Real-Time Detection and Tracking of Pedestrians at Intersections Using a Network of Laserscanners

Daniel Meissner, Stephan Reuter, Klaus Dietmayer
Institute of Measurement, Control, and Microtechnology
Ulm University
Ulm, Germany
daniel.meissner@uni-ulm.de
stephan.reuter@uni-ulm.de

***Abstract*— Accident analysis shows that the majority of accidents with body injuries occur in urban areas and more than 50 percent of those urban accidents happen at intersections. Due to that a major aim of the Ko-PER project, which is part of research initiative Ko-FAS, is to improve safety at intersections by infrastructure based perception. To recognize and track the moving objects, a network of laserscanner sensors observes the intersection and provides a 3D profile of the current scene. By means of the 3D measurements a robust and adaptive Gaussian mixture background model is trained to segment the measurements of dynamic objects and static objects. After the segmentation, the foreground points of each sensor are clustered based on the density of the point clouds and finally pedestrians are classified using dimension features. This paper focuses on tracking of pedestrians, which are the most vulnerable road users. In order to be able to integrate dependencies between the states of the pedestrians, a random finite set particle filter is used to track the pedestrians. The performance of the laserscanner based tracking system is shown and evaluated with measurements from the Ko-PER test intersection at Conti-Safety-Park. Therefore, the optimal subpattern assignment (OSPA) metric is used to evaluate the object recognition and tracking system.**

I. Motivation

Accident scenarios at intersections are amongst the most complex owing to the number and variety of road users, intersection layout, speed range and the different directions from which traffic may approach. In addition, latest German accident analysis shows that the number of injured pedestrians at urban intersections increases [1]. Due to high traffic density and complexity at intersections, pedestrians are most likely to be occluded by other road users or simply missed. To defuse intersections as a black-spot for accidents with pedestrians, a infrastructure based perception system which provides a birds-eye-view of the intersection was developed. A network of multiple laserscanners enables was installed which provides a 3D profile of the current scene and is independent on light changes to observe the intersection and detect moving objects. This contribution focuses on the recognition and tracking of multiple pedestrians. After the intersection system has detected a pedestrian, the estimated dynamic states as well as the dimension and pose parameters of it are communicated to approaching vehicles via wireless Infrastructure-To-Vehicle (I2V) communication.

In recent years, intersection monitoring has attached much attention. Beside video also laserscanner based monitoring systems are presented in literature. In [2] a network of laserscanners, which are mounted on roadsides 0.4 meters above the ground, was used to recognize and track moving objects at intersections. Moving objects are segmented using a static background model and tracking is performed with a Kalman filter and a pedestrian walking model. This system observes a three-way intersection but due to the low mounting position of the sensors the system is pronsegemented to occlusions. A sensor setup similar to the one used in this paper is proposed in [3]. Here the sensors are mounted several meters above the street level. Measurements of objects are classified and used for a contour-based object tracking.

The major contibution of this publication is an adaptive and robust method to recognize pedestrians as well as a real-time random finite set particle filter which considers object interaction in pedestrian tracking.

The publication is structured as follows: After the introduction of the test intersection a pedestrian recognition method is proposed. The method starts with an adaptive background estimation based on a Gaussian mixture model to cope with outdoor conditions like moving vegetation and posts. The resulting foreground measurements are then clustered with an adapted version of the common DBSCAN [4] algorithm. Finally, pedestrians are recognized according to their size. The tracking of the pedestrians using a random finite set particle filter is explained in section IV. Finally, section V shows results of the pedestrian recognition and tracking based on measurements of the test intersection in Conti-Safety-Park. Here the optimal subpattern assignment (OSPA) [5] metric is used to evaluate the performance of object recognition and tracking system. The application of the OSPA metric to evaluate a object recognition system is, to the best knowledge of the authors, a novelty in this field.

II. Ko-PER Test Site at Conti-Safety-Park

A development and test installation of the laserscanner perception system has been built up at the Conti-Safety-Park in Alzenau, Germany. The intersection has been equipped with eight SICK LD-MRS research multilayer laserscanners, which are mounted at a six meter high rack (see Fig. 1 and

978-1-4673-2118-1/$31.00 ©2012 IEEE

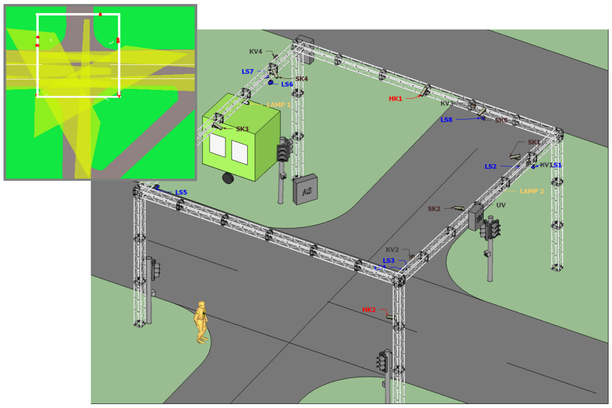

Fig. 1. Sketch of the Ko-PER test intersection at the Conti-Safety-Park and simulation of the field of views (yellow) of the laserscanners. Mounting positions of the sensors are marked with LS1 - LS8.

Fig. 2. Two laserscanners mounted at the rack at Conti-Safety-Park.

Fig. 2). Because of missing infrastructure components like lamp posts and high traffic lights at the intersection the rack had to be used. The laserscanner network observes three accesses to the intersection. The sensors are synchronized in hardware, hence they operate at the same measurement frequency of 12.5 Hz. High precision time stamps to each trigger are generated for the time to measurement association. The measurements of all sensors are transformed to a fixed local coordinate system of the intersection.

III. Moving Object Detection and Classification

In the center of the intersection, the laserscanner network provides a dense 3D profile of the current scene. As a first, step an adaptive background model is estimated. The background contains all measurement points reflected from static objects like road surface or infrastructure components and points reflected from small periodic moving objects like vegetation.

A. Robust and Adaptive Background Model

The used laserscanner scans the environment with a fixed step size along the azimuth angle γ and elevation angle β (Fig. 3). The angular step size in γ is 0.5° and 0.8° in β. Thus, the used laserscanner provides up to 1360 distance measurements per measurement cycle. Each single distance measurement is modeled with an independent mixture of K Gaussian distributions.

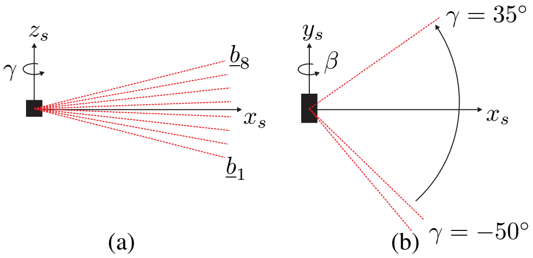

Fig. 3. Vertical (a) and horizontal (b) working range of the laserscanner. Visualization of the measurement steps in γ and β.

Each Gaussian represents a different reflecting point in 3D. The Mixture of Gaussians (MoG) [6] models the probability that a certain measurement has a value of $\underline{x}_t$ at time t and can be written as

$$p(\underline{x}_t) = \sum_{i=1}^{K} \omega_{i,t} \mathcal{N}_x(\underline{\mu}_{i,t}, \Sigma_{i,t}), \tag{1}$$

where ω_i is the weight of the i-th Gaussian component and $\mathcal{N}_x(\mu_{i,t}, \Sigma_{i,t})$ is the corresponding $N = dim(\underline{x})$ dimensional multivariate normal distribution

$$\mathcal{N}_x(\mu_{i,t}, \Sigma_{i,t}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})} \tag{2}$$

with the mean $\underline{\mu}$ and the covariance matrix Σ of $\underline{x}$.

The more measurements result from a static object the more probable the background model is for this measurement. In [6] the proportion $\omega_i / |\Sigma_{x_i}|$ was introduced to decide which portion of the MoG best represents the background measurements. This value increases both as a distribution gains more evidence and as the variance decreases. So the Gaussians are ordered by the value of this fitness score. Finally, the first B Gaussians are used to model the background with

$$B = \arg\min_b \left(\sum_{i=1}^{b} \omega_{i,t} \geq T \right). \tag{3}$$

The parameter T is a threshold which is the minimum prior probability that the background is in the scene [7]. Each measurement which is more than $\delta|\Sigma_i|$ away from any of the B distributions is marked as foreground point. To adapt the background model to actual changes in the scene, the model parameters are updated online. The first Gaussian component that matches

$$\sqrt{(\underline{x} - \underline{\mu}_i)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_i)} \leq \delta|\Sigma_i| \tag{4}$$

is updated with

$$\omega_{t+1} = (1-\alpha)\omega_t + \alpha \tag{5}$$

$$\rho = \alpha \mathcal{N}_x(\underline{x}_{t+1}, \underline{\mu}_t, \Sigma_{i,t}) \tag{6}$$

$$\underline{\mu}_{t+1} = (1-\rho)\underline{\mu}_t + \rho \underline{x}_{t+1} \tag{7}$$

$$\Sigma_{t+1} = (1-\rho)\Sigma_t + \rho(\underline{x}_{t+1} - \underline{\mu}_{t+1})(\underline{x}_{t+1} - \underline{\mu}_{t+1})^T \tag{8}$$

All other Gaussian components receive just an update of their weight with

$$\omega_{t+1} = (1-\alpha)\omega_t. \tag{9}$$

The learning rate $1/\alpha$ is used to regulate the adaption ability of the background model. If no distribution matches the current measurement $\underline{x}_{t+1}$, the least probable component of the MoG is replaced by a new Gaussian distribution with $\mu = \underline{x}_{t+1}$, a low initial weight and high initial variance. To interpret the MoG as a probability according to (1), a normalization of the weights has to be done. After analyzing several measurements and their combination to estimate the background with the MoG, the distance measure figured out to be most relevant measurement. Hence, each Gaussian component models a different distance measure of the laserscanner. Previous to the online background estimation, the model is trained with measurements of an almost empty intersection and a low training rate $1/\alpha = 200$. To keep objects, which for example stopped at a traffic light in the foreground, the learning rate is set to a value of 6300 in online mode.

B. Clustering of Foreground Measurements

After the segmentation of the measurements reflected from moving objects, the measurements representing one object have to be clustered. Therefore the density-based spatial clustering of applications with noise (DBSCAN) algorithm proposed by Ester *et al* [4] is used. The algorithm needs just two parameters and does not require a prior knowledge of the number of clusters. The parameters are the search radius ϵ and the minimum number of points to form a cluster C_{min}. The algorithm starts with an arbitrary point $\underline{p}$ and checks if at least C_{min} points are in the ϵ-neighborhood of it. If this is the case, a cluster is started, $\underline{p}$ is marked as core point and all points in the ϵ-neighborhood belong to this cluster. In the next step, the ϵ-neighborhood of all these points is checked and thus the cluster grows if another core point is found. Points which are in no ϵ-neighborhood and are not surrounded of C_{min} or more points within ϵ distance are labeled as noise. Here, the Euclidian distance is used as distance measure. Obviously the complexity of this clustering algorithm is $O(N^2)$ if N points have to be clustered. To guarantee real-time processing, the algorithm was modified. There are no moving objects above each other, so the clustering in 2D is sufficient. According to that, all foreground points are projected to the x-y-plane of the intersection coordinate system. To reduce the number of distance calculations the x-y-plane is partitioned into quadratic grid cells with edge length ϵ. Instead of calculating the distance of the current examined foreground point to every other point, the proposed method reduces the candidates on those points which are inside the neighboring grid cells (Fig. 4). Especially in the present application area the points appear concentrated in the area of objects, so the partition into grid cells reduces the complexity noticeable without losing accuracy.

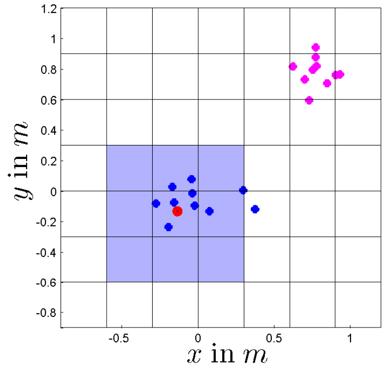

Fig. 4. DBSCAN speed: Projection of the 3D measurement points to a 2D grid. The examined point is marked red and the search area is marked light blue.

C. Pedestrian Recognition

To decide if the cluster of measurement points represents a pedestrian, the dimension of the cluster is analyzed. Therefore, the bounding box of each point cloud is calculated using principal component analysis. With a transformation of the points of the cluster according to the principal components, the dimensions along the x-, y- and z-axes (length, width and height) can be calculated as well as the orientation of the box around its mean value. The current classification is just based on static properties of the clusters. Analysis of manually labeled sequences pointed out that the length, width and principal component of the cluster are a good feature. This publication focuses on the pedestrian detection at the center of the intersection. Here, clusters with a dimension lower than one meter in length and width as well as the major direction of the principal component is pointing in z-direction are classified as pedestrians and tracked with a random finite set filter.

IV. Multi-Object Tracking

The multi-object Bayes filter [8] is an extension of the single-object Bayes filter. By using a random finite set $\mathbf{X}$ of state vectors, a state of the filter describes the complete environment. Thus, the multi-object Bayes filter is called random finite set (RFS) filter in the following. This section summarizes some details of the multi-object Bayes filter. For more details about the multi-object Bayes filter, refer to [8], [9].

The RFS Bayes filter is implemented using Sequential Monte Carlo (SMC) methods [9], [10], [8]. The difference to well known SMC implementations of the single-object Bayes filter is that a particle set, which represents a random finite set, is used instead of a particle which represents a state vector. Thus, each particle set holds several particles and the number of particles depends on the number of objects in the scene. Further, the number of particles in a particle set may change from one time step to the next.

Since a state of the RFS filter represents the whole environment, the Markov model used in the predictor step has to incorporate object appearance and disappearance in addition to the motion of the objects. In our implementation, the Markov model only represents motion and disappearance, since a measurement driven birth model is used which is evaluated separately. The birth model initializes a new object with a birth probability of $p_B = 0.01$ for each measurement which had no particles in its 3σ range during the calculation of the previous corrector step.

Due to the set representation it is possible to incorporate constraints about possible states of a pedestrian. Since the pedestrians are represented as point targets in the filter, there has to be a minimum distance between any of the pedestrians due to their minimum size. Thus, as proposed in [9] the weight of each particle which contains state vectors which are too close together is decreased.

In the corrector step, the evaluation of the multi-object likelihood function $f_{k+1}(\mathbf{Z}_{k+1}|\mathbf{X}_{k+1|k})$ is necessary. The multi-object likelihood functions for different scenarios (e.g. missed detections and no false alarms) are given in [8]. These likelihoods are based on the assumption that a measurement is generated by no more than one object. In [11] a real-time approximation of the multi-object likelihood using a graphical processing unit (GPU) is proposed. By allowing to associate a measurement to more than one object, the computational complexity drastically decreases. Under the assumption that the minimum distance between two objects in the set is approximately as large as $6\sigma_z$, where σ_z is the standard deviation of the measurement noise, the value of the approximated multi-object likelihood is very close to the exact multi-object likelihood.

Since the RFS filter only estimates the multi-object posterior density function and the number of objects in the scene, a subsequent extraction of the individual objects is necessary. To simplify the track extraction, a track label l_i is appended to each state vector. Thus, it is possible to initialize the track extraction algorithm with the labeling result of the previous step. An adapted k-means [12] algorithm, which estimates the number of cluster centers itself, is used to extract the tracks. Only new-born particles, which are more than the minimum size of a pedestrian away from any of the cluster centroids, are used to initialize an additional centroid. On the other hand, if no particle is assigned to one of the centroids any more, the centroid is removed.

V. Results

Beside the visual evaluation of the pedestrian recognition method the OSPA metric [5] is used to evaluate the performance of the clustering algorithm and the tracking system. The OSPA distance incorporates the spatial as well as the cardinality error. The OSPA distance represents the affinity of two sets using just a single distance value. If the number of objects is equal in both sets and if they are located exactly at the same position, the OSPA distance is $d_{OSPA} = 0$. The maximum value $d_{OSPA} = c$, where c is the cut-off parameter, is reached, if e.g. one of the sets is empty while the other one is not empty. For the evaluation of the tracking system, which processes the pedestrian hypothesis of the clustering algorithm, a challenging sequence with up to ten pedestrians in the center part of the intersection is used. As can be seen by means of the ground truth trajectories in Fig. 5, the pedestrians move unregulated, cross each other and move close together.

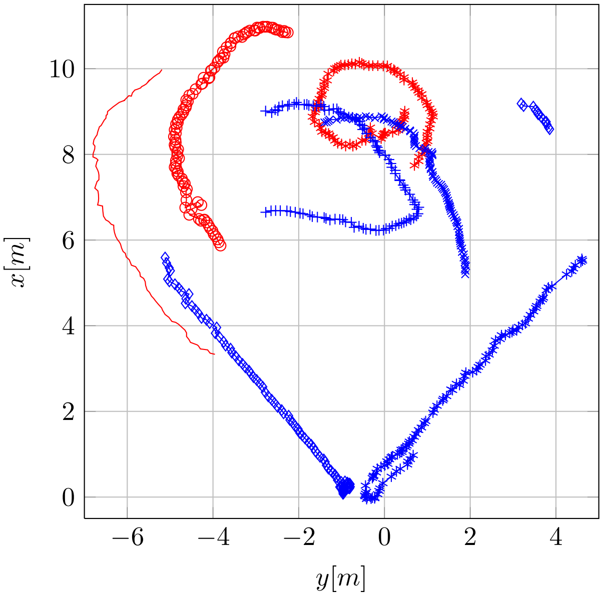

Fig. 5. Ground truth trajectories of all pedestrians for time steps $k = 200$ to $k = 300$.

A. Hypothesis Generation

The performance of the background estimation and the pedestrian recognition is depicted in Fig. 6 and Fig. 7 respectively. In Fig. 6, the measurement points of all eight laserscanners at the test intersection in Alzenau are shown. As can be seen with the help of the different colors, the measurements resulting of background objects are classified reliably. The measurement points reflected from moving pedestrians are depicted in black. In Fig. 7 (a) and (b), clustered and classified foreground measurements are plotted. As mentioned in section III, the clusters of pedestrians differ significantly in their dimension. Thus, the proposed classification features provide good results in the center part of the intersection, where the point density is very high.

So far, the OSPA metric was mainly used in the evaluation of multi-object tracking algorithms. Due to its characteristics, the OSPA metric is perfectly suited to evaluate the performance of clustering algorithms, too. In our evaluations, the cut-off parameter of the OSPA distance is chosen to $c = 10$ and the order of the metric is set to $p = 1$. Fig. 8 shows the value of the OSPA distance.

The OSPA distance confirms the impression of the visual inspection. Just in the part of the sequence where two or more pedestrians are very close to each other the clustering algorithm is not able to separate them from each other. That is because the DBSCAN algorithm has fixed values for the ϵ-neighborhood and C_{min}. The effect can be observed at $k = 270$ and $k = 310$. Here, the OSPA value shows a

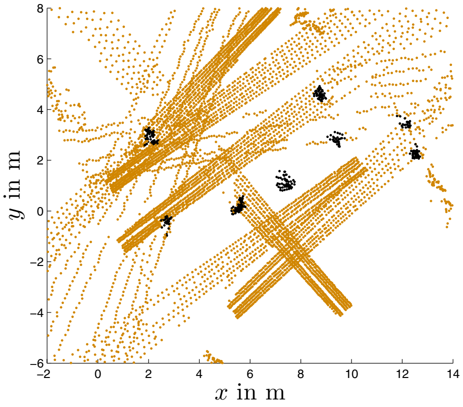

Fig. 6. Foreground segmentation. Foreground points are marked black, background orange.

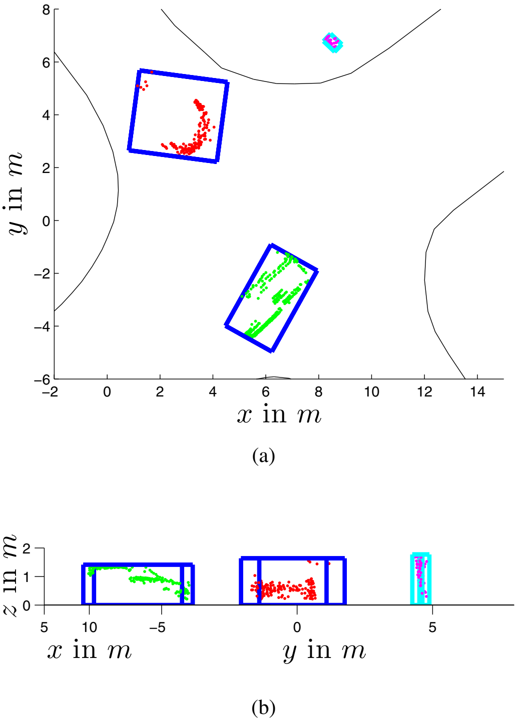

Fig. 7. Clustered and classified foreground points. Different clusters are shown in different colors. Cyan bounding boxes mark pedestrians, blue boxes mark vehicles. Roadside is shown in black.

step of about one, which is an additional penalty for the cardinality error which results from merging two pedestrians into one cluster. Hence, a second clustering step which incorporates the characteristic shape of pedestrians in the laser measurements to separate those objects is currently under development.

B. Pedestrian Tracking

The results of the clustering algorithm are finally used to track all persons at the intersection. Like the clustering algorithm, the RFS filter is also evaluated using the OSPA metric.

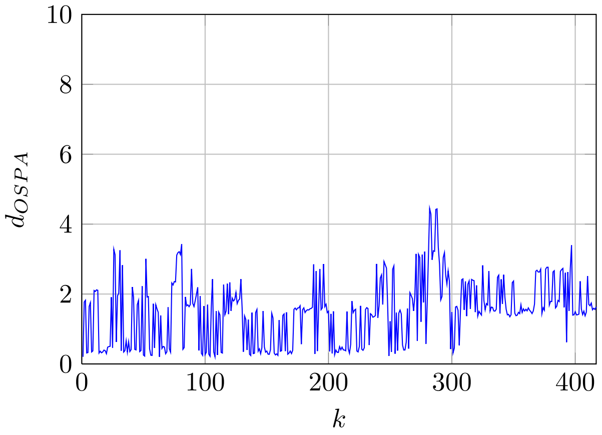

Fig. 8. OSPA distance of the clustering.

Fig. 9 shows the mean value of the OSPA distance for 50 Monte Carlo runs. In comparison to the OSPA distance of the clustering algorithm the distance of the tracking is much smoother. That is because the tracking system bridges the frames where an objects is not detected. The small peaks in the plot of the OSPA distance are due to cardinality errors during object appearance and disappearance. On the one hand, the birth model leads to a delay of the appearance of at least one measurement cycle. On the other hand, the track of a pedestrian may persist for several measurement cycles after the last measurement was received due to the persistence probability.

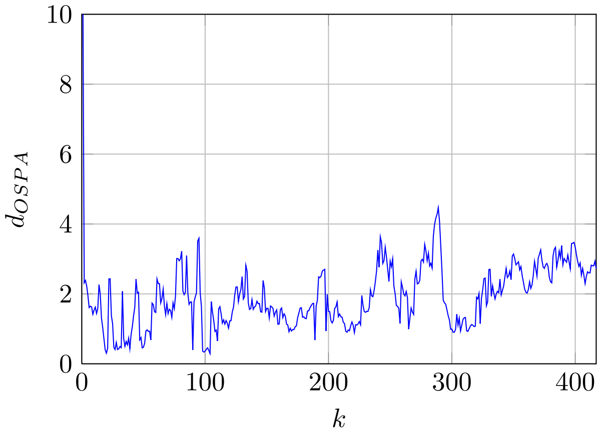

Fig. 9. Mean value of the OSPA distance for 50 Monte Carlo runs of the RFS filter.

Fig. 10 shows the labeled ground truth trajectories as well as the estimated tracks using the RFS filter. The ground truth values and the estimates are very close to each other most of the time. In the upper middle, where the trajectories of three pedestrians cross several times, it is possible that estimated tracks jump from one ground truth trajectory to another. This is due to the fact, that the clustering algorithm is not able to separate two persons for several time steps.

The RFS filter is implemented using CUDA [13] on a Nvidia Tesla C2075 GPU. The average computing time of

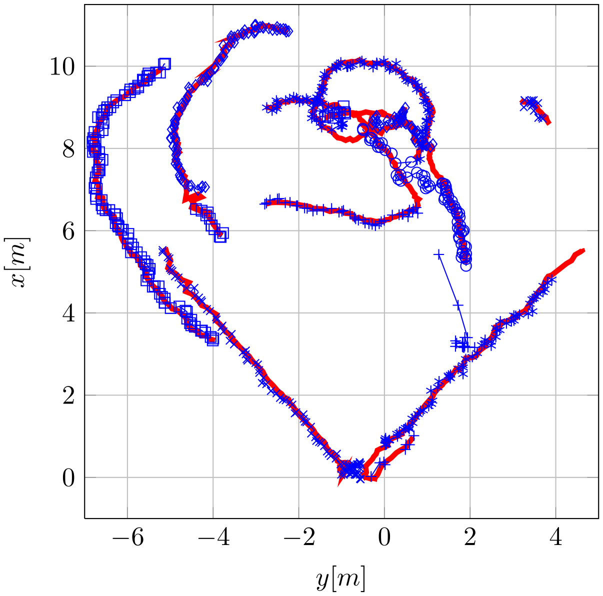

Fig. 10. Ground truth trajectories and estimated trajectories of the RFS filter for all pedestrians of time steps $k = 200$ to $k = 300$. The ground truth is marked by a thick red line while the estimated trajectories are marked by blue lines with additional markers.

the RFS filter with $N = 25000$ particle sets is 35.87ms. This time includes the calculation of the predictor and corrector step as well as the track extraction algorithm. Further, all data transfers between CPU and GPU to upload the measurements and to download the extracted tracks are included. Since the measurement rate of the sensors is 12.5 Hz, real-time processing is possible.

VI. Conclusion and Future Works

Infrastructure based perception systems can increase the traffic safety considerably because they monitor even complex urban intersections. In this publication, a pedestrian recognition and tracking system has been presented. The proposed Gaussian mixture background segmentation method estimates the background of the current scene. Doing so, one is able to adapt the background online to changes at the intersection and compensate repetitive motions in the background, which is necessary in outdoor applications. After the segmentation of the measurements resulting from moving objects, they are clustered using the DBSCAN algorithm. The advantage of being independent of the number of clusters is exploited and the drawback in computation complexity is reduced. In the last step of the pedestrian recognition, the clusters are classified by means of dimension features.

Finally, the recognized pedestrians are tracked with a random finite set particle filter. The set representation allows to integrate dependencies between the possible states of a pedestrian. This becomes noticeable in the robustness of the pedestrian tracks. The pedestrian recognition and tracking system is evaluated with a challenging sequence with up to ten pedestrians. The OSPA distance over the sequence verifies an accurate tracking performance.

Additionally, the evaluation of clustering algorithms using the OSPA metric was proposed. Using the OSPA metric, parameters of the clustering algorithm can be easily optimized and the performance of different clustering algorithms can be compared.

Further steps might refine the clustering and classification method. For example the shape of pedestrians, vehicles and other road users might be introduced to enable the separation of objects which are close together. In addition, the tracking system is going to be enhanced by other road users.

VII. Acknowledgments

This work results from the joint project Ko-PER - which is part of the project initiative Ko-FAS - and has been funded by the German *Bundesministerium für Wirtschaft und Technologie* (Federal Ministry of Economics and Technology) under grant number 19S9022G. A part of this work is done within the Transregional Collaborative Research Center SFB/TRR 62 ”Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

References

[1] Destatis, “Press release no. 462,” 2011. [Online]. Available: http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Presse/pm/2011/12/PD11__462__46241.psml

[2] H. Zhao, J. Cui, H. Zha, K. Katabira, X. Shao, and R. Shibasaki, “Sensing an intersection using a network of laser scanners and video cameras,” *Intelligent Transportation Systems Magazine, IEEE*, vol. 1, pp. 31 –37, 2009.

[3] C. Stimming and B. Roessler, “Cooperative infrastructure safety systems using infrastructure laserscanner,” in *Proceedings of the International Workshop on Intelligent Transportation (WIT)*, 2011, pp. 97–102.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” *International Conference on Knowledge Discovery and Data Mining*, pp. 22–231, 1996.

[5] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, “A consistent metric for performance evaluation of multi-object filters,” *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447 –3457, 8 2008.

[6] C. Stauffer and W. Grimson, “Adaptive background mixture models for real-time tracking,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 637–663.

[7] P. Kaewtrakulpong and R. Bowden, “An improved adaptive background mixture model for realtime tracking with shadow detection,” in *European Workshop on Advanced Video Based Surveillance Systems*, 2001.

[8] R. P. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Artech House Inc., Norwood, 2007.

[9] S. Reuter and K. Dietmayer, “Pedestrian tracking using random finite sets,” in *Proceedings of the 14th International Conference on Information Fusion*, 2011.

[10] H. Sidenbladh and S.-L. Wirkander, “Tracking random sets of vehicles in terrain,” in *Conference on Computer Vision and Pattern Recognition Workshop*, 2003.

[11] S. Reuter, K. Dietmayer, and S. Handrich, “Real-time implementation of a random finite set particle filter,” in *Sensor Data Fusion: Trends, Solutions, Applications (SDF 2011)*, 2011. [Online]. Available: http://www.user.tu-berlin.de/komm/CD/paper/100149.pdf

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.

[13] Nvidia, “Nvidia cuda toolkit 4.0,” 2011. [Online]. Available: http://developer.nvidia.com/cuda-toolkit-40