CONFIDENCE-AWARE PEDESTRIAN TRACKING USING A STEREO CAMERA

U. Nguyen*, F. Rottensteiner, C. Heipke

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
(nguyen, rottensteiner, heipke)@ipi.uni-hannover.de

Commission II, WG II/4

KEY WORDS: pedestrian tracking, stereo camera, tracking-confirm-detection, detection confidence, trajectory confidence

ABSTRACT:

Pedestrian tracking is a significant problem in autonomous driving. The majority of studies carries out tracking in the image domain, which is not sufficient for many realistic applications like path planning, collision avoidance, and autonomous navigation. In this study, we address pedestrian tracking using stereo images and tracking-by-detection. Our framework comes in three primary phases: (1) people are detected in image space by the mask R-CNN detector and their positions in 3D-space are computed using stereo information; (2) corresponding detections are assigned to each other across consecutive frames based on visual characteristics and 3D geometry; and (3) the current positions of pedestrians are corrected using their previous states using an extended Kalman filter. We use our tracking-to-confirm-detection method, in which detections are treated differently depending on their confidence metrics. To obtain a high recall value while keeping a low number of false positives. While existing methods consider all target trajectories have equal accuracy, we estimate a confidence value for each trajectory at every epoch. Thus, depending on their confidence values, the targets can have different contributions to the whole tracking system. The performance of our approach is evaluated using the Kitti benchmark dataset. It shows promising results comparable to those of other state-of-the-art methods.

1. INTRODUCTION

Image-based multiple objects tracking is a critical problem in the fields of computer vision and robotics. Pedestrians are one of the most relevant objects to be tracked, motivated among others by the development of applications related to autonomous driving and traffic safety. Tracking allows vehicles not only to know where pedestrians appear, but also to anticipate their moving directions and behaviors, which are crucial factors for planning their driving paths and safe navigation.

Despite recent advances, the performance of existing trackers still needs to be improved significantly to close the gap between human and machine perception performance, so that computer systems can assist or fully replace human efforts on practical tasks (Leal-Taixé et al., 2017). The tracking-by-detection paradigm is used by most multi object tracking systems (Henschel et al., 2018; Linder et al., 2016; Yoon et al., 2015). This approach first detects target objects in each image independently, then corresponding detections are associated w.r.t. each other across frames. The recent emergence of convolutional neural networks (CNNs) resulted in many powerful detectors (He et al., 2017, 2016; Zhang et al., 2016); however, they still have the problem of increasing the number of false positives (FPs) together with the recall. Hence, we aim at obtaining a high number of true positives (TPs), but still keep FP at a low rate. We do so modifying the association step of the tracking pipeline, which connects results of consecutive frames: in this step, while employing all detections of the current frame as input for the assignment, we use solely highly accurately detected pedestrians to create a new trajectories, a strategy called tracking-to-confirm-detection (TCD). We also estimate a confidence value for each trajectory, and we recover missed detections, e.g. due to occlusions, during tracking, by employing tracklet extrapolation. However, while facilitating the increase of TP, the extrapolation can accumulate more false alarms as well. To reduce this negative side of the extrapolation, we keep tracked targets with high confidence values longer in the system than the weak ones.

While several studies have focused on tracking interesting objects in image space (Breitenstein et al., 2011; Fagot-Bouquet et al., 2016; Kieritz et al., 2016; Leal-Taixé et al., 2017), automobiles require 3D location and trajectory information of pedestrians in object space. Using monocular image sequences, it is challenging to predict and localize objects in world coordinates due to the small baselines associated with near real-time requirements. To overcome this problem, we develop a tracking-by-detection approach using stereo images, which makes it possible to estimate 3D positions of tracked pedestrians. As the quality of the 3D information derived from stereo images depends on the baseline, the distance from an object to the stereo system and the quality of the matching algorithm, we combine both 2D and 3D information to track people more accurately. We also correct the velocity of each pedestrian in object space based on its neighbors using our motion model, in which we incorporate our trajectory confidence value. In order to demonstrate the competitive performance of our tracker, we conduct the experiments on the Kitti tracking data set (Geiger et al., 2012). The results are analyzed and compared with other state-of-the-art methods.

Our main contributions can be summarized as follows:

1. We introduce a framework to track pedestrians by employing both 2D and 3D information. Stereo images are used to model the scene and estimate pedestrian positions in 3D object space. The appearances of pedestrians in image space are utilized for detection and spatio-temporal feature comparison.

*Corresponding author

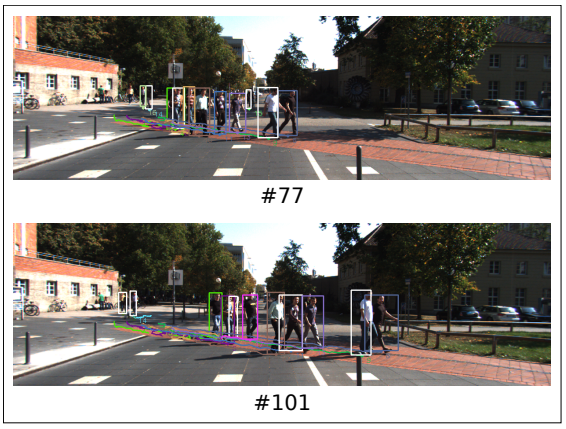

Figure 1. An exemplary tracking results of our tracker. The generated pedestrian trajectories are back-projected to image space at two different epochs.

2. We suggest the TCD approach to obtain high recall and small false alarm values of detections during tracking.

3. We propose considering detections and trajectories differently, depending on their confidence values. Additionally, we develop a motion model to correct the estimated movements of tracked objects utilizing their highly accurate neighbors.

The rest of this paper is organized as follows: in Section 2, we discuss previous studies related to our research. We describe the details of our tracking framework in Section 3. The performance of our tracker is presented in Section 4, followed by the conclusion in Section 5.

2. RELATED WORK

Multi-people tracking: Most of the modern trackers employ the tracking-by-detection approach to continuously localize and identify pedestrians in image sequences (Choi, 2015; Dehghan et al., 2015; Henschel et al., 2018; Hong Yoon et al., 2016; Klinger et al., 2017; Pirsiavash et al., 2011; Zamir et al., 2012). This method usually comes in three phases: (1) pedestrians are detected in each image; (2) detections in consecutive frames are associated into consistent sets of trajectories; and (3) a filter step is performed to smooth the trajectories based on their previous states. The core of this approach is the data association step, which is based primarily on visual and geometry cues.

In general, data association is carried out either as a local (online) method or as a global association. For the online approach, the pedestrians are linked across frames in a pairwise fashion. Since only detections of two frames are considered, this method is vulnerable to wrong detections (Breitenstein et al., 2011; Choi, 2015; Fagot-Bouquet et al., 2016; Kieritz et al., 2016; Lenz et al., 2015; Xiang et al., 2015). Global methods, on the other hand, generate tracklets or complete trajectories from a batch of frames or the whole image sequence. This enables global properties of target objects to be taken into account during the optimization. That is why most global matchers usually outperform the local approaches (Berclaz et al., 2011; Dehghan et al., 2015; Pirsiavash et al., 2011; Zamir et al., 2012; Zhang et al., 2008). Nevertheless, requiring the entire image sequence before performing tracking, global techniques can only be used for offline cases. In applications where an instant response is a significant demand, like for autonomous driving or robot-human interaction, only online approaches are appropriate.

While most of state-of-the-art methods execute tracking in 2D image space and concentrate on correcting the assignments, positions and moving directions of pedestrians in 3D object space are essential prerequisites for vehicles to automatically manage their motions. For this reason, several systems do tracking based on stereo or RGB-D cameras or sensors based on structured light. Although widely used for indoor tracking studies (Jafari et al., 2014; Linder et al., 2016), RGB-D devices are not appropriate for outdoor environment due to illumination problems and complicated surfaces. Some publications (Mitzel et al., 2010; Ošep et al., 2017; Schindler et al., 2010) proposed using a stereo rig, mounted on a mobile platform to track people on streets. The 3D geometric position of a pedestrian is estimated by inspecting the detected bounding box or intersecting the image space detection with the ground plane. Estimating the foot positions of pedestrians on the ground plane allows reducing pedestrians movement in 3D-space from three dimensions to two dimensions, as they are supposed to walk on the road.

Motivated by autonomous driving applications, we carry out pedestrian tracking in 3D-space using stereo images and follow the tracking-by-detection approach. We apply bipartite matching to associate interesting objects in adjacent frames. However, instead of using only information of two contiguous epochs that might contain high uncertainties and errors, we aggregate information from a certain number of previous epochs to increase the accuracy of data association. In addition, we compute confidence scores for the trajectories in each frame, which can help to improve the matching and tracking precision.

Motion model: To produce a reliable trajectory over time, the state of a pedestrian predicted from its previous positions can be exploited to correct its current state. For the prediction, various motion models were proposed, in which the movement of a person is influenced by other people nearby. Zhang and van der Maaten (2013) suggested predicting the position of a pedestrian by observing the movements of its neighbors. Similarly, also applying a grouping model, Klinger et al. (2017) improved this method by weighting the effect of each neighbor based on an angular displacement of its moving directions compared to the current person. Yoon et al. (2015) and Leal-Taixé et al. (2014) proposed anticipating the states of a target based on the history of all observed trajectories, where the movement of irrelevant people, which might affect the results.

Adopting the explicit grouping approach, we perform movement prediction of pedestrians and consider their interactions with people nearby. However, different from previous studies, the impact of a neighbor on a certain person is determined by their spacial distance and moving direction difference. In addition, while using neighbors with highly accurate trajectories can improve the prediction reliability, including those with low confidence values into the motion model can lead to the accumulation of incorrect information. Therefore, only trajectories with high confidence are considered as candidates in our motion model.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W5-53-2019 | © Authors 2019. CC BY 4.0 License.

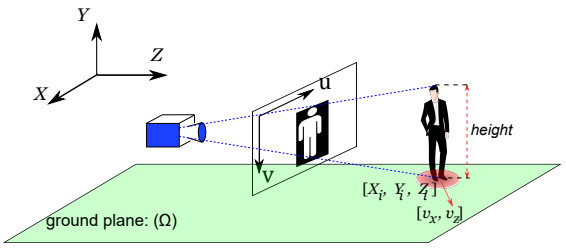

Figure 2. Pedestrian localization in 3D space by intersecting the detection with the ground plane. The stereo pair is embedded for the sake of clarity.

3. METHODOLOGY

Aiming at tracking pedestrians in 3D-space, our tracker takes normalized stereo image pairs as input and estimates trajectories of observed pedestrians in a common 3D coordinate system with directions of axes as shown in Figure 2. Following the tracking-by-detection technique, our framework consists of three primary phases: detection and post-processing, data association, and prediction and filtering. Figure 3 depicts an overview of the processing chain of our tracking framework. The details of each step are described in the subsequent sections.

3.1 Detection and post-processing

Scene modeling: Given a calibrated stereo image pair, the disparity map w.r.t the stereo rig is estimated using the state-of-the-art dense matching approach presented in (Yamaguchi et al., 2014). Afterward, a 3D point cloud $\mathcal{P}$ is computed from the disparity map via stereo triangulation. We assume that a scene primarily consists of vertical planar objects, e.g. building facades and pedestrians, supported by the horizontal ground plane (e.g. the road). We follow the approach presented by Nguyen et al. (2018) to model the scene (see Figure 4), generating the following pieced information:

- an obstacle mask $\mathcal{M}_o$ mainly corresponding to building facades;
- an area of interest mask $\mathcal{M}_{in}$, indicating areas where pedestrians can appear in the image;
- 3D ground plane (Ω) in object space.

Person detection and localization: We adopt the pre-trained mask R-CNN method (He et al., 2017) to detect people in images. For each detected object, mask R-CNN provides:

- the upper left corner, width, and height of a 2D bounding box (BB): $bb = \{r, c, w, h\}$, which covers the area where the object of interest exists in the input image;
- its confidence ϱ about the classified type of that object;
- a binary mask $\mathcal{M}_{seg}$ to separate foreground and background in each BB.

Beside the high accuracy, the instance segmentation mask $\mathcal{M}_{seg}$ is a big advantage of mask R-CNN. This mask simplifies the estimation of the position and height of a target in object space. All detections classified as humans and having a confidence value ϱ larger than a threshold $\epsilon_{\varrho 1}$ are considered for post-processing.

To localize an object (pedestrian) in 3D, we project all 3D points belonging to that object in $\mathcal{M}_{seg}$ to the ground plane (Ω) and average them to obtain the foot point $P^{\mathsf{F}} = [X^{\mathsf{F}}, Y^{\mathsf{F}}, Z^{\mathsf{F}}]$ of the object. The positions of the foot point in images $M = [u, v, d]$ are estimated by back-projecting P^{F} into images, where u and v are image coordinates in the left image and d is the disparity value. This procedure often allows us to compute the 3D position and recover the entire body of an observed object in the input image even if only parts are visible (see Figure 5). The uncertainty $\sigma_M = [\sigma_u, \sigma_v, \sigma_d]$ of M is heuristically estimated. σ_u and σ_v are fixed, and σ_d is determined based on the accuracy of matching algorithm (Yamaguchi et al., 2014). The uncertainty of the position in 3D σ_P is then computed through error propagation.

We assume that points in the mask $\mathcal{M}_{seg}$ and have smallest v value are head points of a detected object. Employ those head points in images and the point cloud $\mathcal{P}$, we also estimate the head position of interesting objects in 3D: $P^{\mathsf{H}} = [X^{\mathsf{H}}, Y^{\mathsf{H}}, Z^{\mathsf{H}}]$, which we use together with the foot point position to compute the object heights: $height = Y^{\mathsf{H}} - Y^{\mathsf{F}}$.

Mask R-CNN is only based on image visual information to detect persons and thus yields a number of false alarms. These can partly be detected and eliminated by utilizing additional 3D properties as follows:

- Pedestrian heights ($height$) are limited in a certain range.
- Pedestrians must appear in the area of interest: $bb = bb \cap \mathcal{M}_{in}$.
- A pedestrian should not completely lie inside obstacle mask: $bb \neq bb \cap \mathcal{M}_o$.

Detected objects that do not satisfy these three constraints are not further considered in the tracking phase.

3.2 Data association

System setup: Let $\mathcal{D} = \{D_{1,t}, ..., D_{n,t}\}$ and $\mathcal{T} = \{\tau_{1,t}, ..., \tau_{m,t}\}$ be n observations and m target trajectories at time t, respectively. Each observation $D_{i,t}$ includes its positions of the foot point in both the stereo images $M = [u, v, d]$ and 3D-space $P = [X, Y, Z]$, the corresponding uncertainties, the detection confidence ϱ, and the 2D bounding box bb:

$$D_{i,t} = \{M, \sigma_M, P, \sigma_P, bb, \varrho\} \,. \tag{1}$$

The trajectories $\tau_{j,t} = \{S_{j,k}, ..., S_{j,t-1}\}$ contain the state history of a tracked person up to epoch $(t-1)$, in which a state $S_{j,k} = [X, Y, Z, v_x, v_z]^T$ consists of 3D position and velocity. The uncertainty Σ_{SS} of a state is estimated by the extended Kalman filter (see Equation (9)). In the 3D coordinate system, pedestrians are assumed to move on the ground plane, so there is no movement in Y direction.

A trajectory target is considered to be deactivated if it is not assigned to any observation and becomes activate again if there is a detection assignment in the future. Positions of a deactivated target are still predicted for a number of epochs until that target is completely deleted in the tracking system.

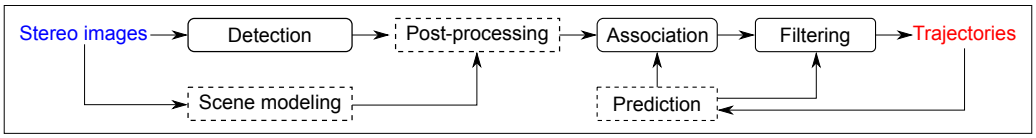

Figure 3. Our general tracking framework.

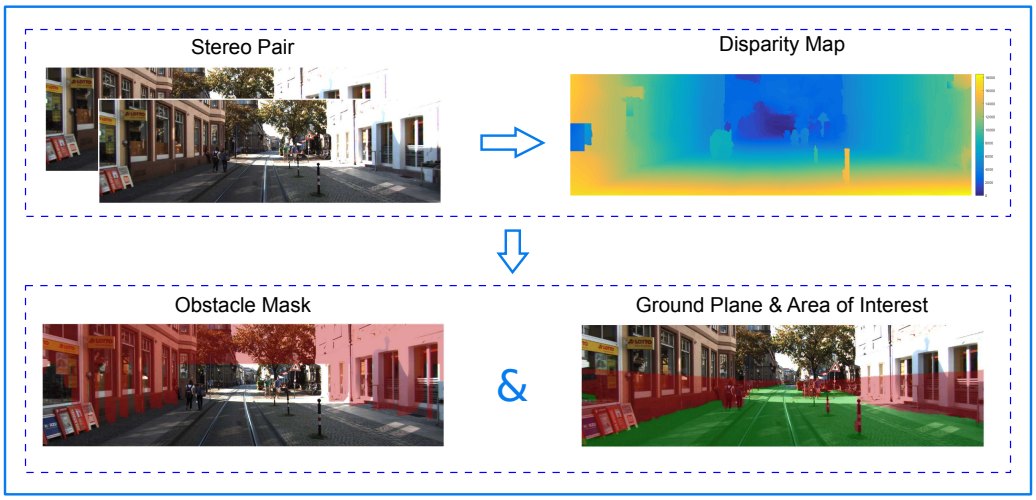

Figure 4. Scene modeling. Using stereo information, we generate three different areas in the image space including obstacle, ground plane and area of interest.

(a) (b)

Figure 5. The detected 2D bounding box (a) is corrected using the back-projected foot point from 3D (b).

We optimize the problem of assigning detections in $\mathcal{D}$ to targets in $\mathcal{T}$ using a binary integer program described as follows:

$$\begin{cases} \text{maximize} & c^T w \\ \text{subject to} & (Ac)_k \leq 1, k = 0, ..., (n+m) \end{cases}, \quad (2)$$

where $c = \{c_i^j, ..., c_n^m\}$ is an indicator vector. For $c_i^j = 1$ the detection $D_i \in \mathcal{D}$ and trajectory $\tau_j \in \mathcal{T}$ are associated with each other, otherwise $c_i^j = 0$; The association weight $w_i^j \in w = \{w_i^j, ..., w_n^m\}$ describes how likely D_i and τ_j belong to one and the same person; A is a $(n+m) \times (nm)$ design matrix and has the effect that one detection is assigned to at most one trajectory and vice versa.

Association weight: This weight describes the likelihood that an observation is assigned to a target, which is primarily explained by its visual $\Gamma_{\mathcal{A}}$ and spatial distance $\Gamma_{\mathcal{G}}$ similarity. Beside that, a high confidence detection is preferred to be allocated to existing trajectories over one with low confidence. In the same manner, a trajectory with a high confidence $\vartheta_{\tau_{j,t}}$ is more likely to continue to be observed in the current frame.

Our association weight is computed as follows:

$$w_i^j = \rho\Gamma_{\mathcal{G}}(D_{i,t}, \tau_{j,t}) + \theta\Gamma_{\mathcal{A}}(D_{i,t}, \tau_{j,t}) + \nu\varrho_{D_{i,t}} + \iota\vartheta_{\tau_{j,t}}, \quad (3)$$

where ρ, θ, ν, and ι are parameters used to define the impact of each criterion on the association weight value. The component $\Gamma_{\mathcal{G}}$, $\Gamma_{\mathcal{A}}$, and $\vartheta_{\tau_{j,t}}$ are defined in the following paragraphs.

Geometry similarity: this value is related to the 3D spatial distance of an object and its potential target. Let $S_{j,t}^+$ is a predicted state of $\tau_{j,t}$ at an epoch t, which is estimated by the Kalman filter (see Equation (7)). We compute the Mahalanobis distance in 3D space between the predicted position at t of $\tau_{j,t}$ and the position of $D_{i,t}$ as their geometry affinity and this distance is mapped to a value in the range of from 0 to 1 by an exponential function to obtain the criteria $\Gamma_{\mathcal{G}}$:

$$\begin{aligned} \phi_{\mathcal{G}}(D_{i,t}, \tau_{j,t}) &= (S_{j,t}^+ - P_{D_{i,t}})^T \Sigma_{SS,t}^+ (S_{j,t}^+ - P_{D_{i,t}}) \\ \Gamma_{\mathcal{G}}(D_{i,t}, \tau_{j,t}) &= e^{\frac{\phi_{\mathcal{G}}(D_{i,t}, \tau_{j,t})}{-\varepsilon_{\mathcal{G}}}} \end{aligned}, \quad (4)$$

where $\varepsilon_{\mathcal{G}}$ is a free parameter and $\Sigma_{SS,t}^+$ is the predicted variance of $S_{j,t}^+$ (see Equation (7)). In the above calculations, we only use position entries $[X, Y, Z]$ of $S_{j,t}^+$ while the velocity elements are disregarded.

Appearance similarity: The appearance similarity accounts for the resemblance between two objects in image space in terms of texture, color, shape, etc. Beside the geometric similarity, this is a significant cue to distinguish between different persons. The visual properties of a detection are represented by a feature vector f, extracted by TriNet (Hermans et al., 2017). At time t, the feature vector of a trajectory $\tau_{j,t}$ is the average of its appearance vectors from a certain number of previous epochs, which can account for visual properties of a trajectory within a temporal window. The appearance similarity

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W5-53-2019 | © Authors 2019. CC BY 4.0 License.

$\Gamma_{\mathcal{A}}$ between $D_{i,t}$ and τ_j is computed as:

$$\begin{aligned} \phi_{\mathcal{A}}(D_{i,t}, \tau_{j,t}) &= \|f_{\tau_{j,t}} - f_{D_{i,t}}\|_{L_2} \\ \Gamma_{\mathcal{A}}(D_{i,t}, \tau_{j,t}) &= e^{\frac{\phi_{\mathcal{A}}(D_{i,t}, \tau_{j,t})}{-\varepsilon_{\mathcal{A}}}} \end{aligned} , \quad (5)$$

where $\varepsilon_{\mathcal{A}}$ is a free parameter.

Trajectory confidence: We define a confidence value ϑ to represent the accuracy and reliability of a trajectory at a specific epoch. While the accuracy accounts for the possibility of a trajectory to be generated from TP detections of an identical person, the reliability describes how long the trajectory already exists in the system as an active one. These cues are combined to estimate ϑ as follows:

$$\vartheta_{\tau_{j,t}} = \frac{1}{k+1} \sum_{l=t-k}^{t} (\alpha \varrho_{\tau_{j,l}} + \beta w_{\tau_{j,l}}) + \gamma \min(1, \frac{a_{\tau_{j,t}}}{\epsilon_a}) , \quad (6)$$

where k is the number of epochs in the past before t; $\varrho_{\tau_{j,l}}$ is the detection confidence of the observation assigned to τ_j at l with association weight $w_{\tau_{j,l}}$; $a_{\tau_{j,t}}$ is the number of active states that τ_j has until t, which is normalized by a threshold ϵ_a; and α, β, and γ are weight parameters, which define the contribution of each cue on the trajectory confidence value.

Association gate: Since people walk with limited speed, the covered distance in a small amount of time cannot exceed a threshold. Exploiting this property, we generate two geometric gates, which indicate whether a detection can be assigned to a target or not. While the first gate constrains the distance in 3D space between a detected pedestrian and a trajectory, the second gate restricts their overlap area in image space. These gates help to reduce both the complication of the optimization problem and inconsistent assignments.

Tracking-confirm-detection: Since detected pedestrian results are noisy, using a single detection confidence threshold (DCT) is usually hard to achieve high recall and low false alarm at the same time. Therefore, in our tracking-confirm-detection (TCD) approach, we use two predefined DCTs: a low $\epsilon_{\varrho 1}$ and a high $\epsilon_{\varrho 2}$. All detections with a confidence value larger than $\epsilon_{\varrho 1}$ are considered during assignment optimization. The reason for this is because a trajectory can be used to confirm the presence of a TP detection nearby even its confidence value is very low. However, when a new trajectory is created, there is no additional evidence to confirm its correctness other than its detection confidence. Hence, at a specific epoch, a detection which is not assigned to any existing target initializes a new trajectory if its confidence value is larger than $\epsilon_{\varrho 2}$.

3.3 Prediction and filtering

As a trajectory evolves over time, pedestrian states consisting of positions and velocities close in time are correlated. Therefore, the state of the trajectory at a specific epoch can be predicted from its previous states. This predicted state is employed to correct the current measurement using an extended Kalman filter (Gelb, 1974) as follows:

- Let $S_t = [X_t, Y_t, Z_t, v_{X,t}, v_{Z,t}]^T$, $\Sigma_{SS,t}$ be the state and covariance matrix of a target trajectory T_j at t. Its predicted state S^+_{t+1} in the next epoch is calculated through the transition matrix ψ.

- State prediction:

$$\begin{aligned} S^+_{t+1} &= \psi S_t \\ \Sigma^+_{SS,t+1} &= \psi \Sigma_{SS,t} \psi^T + Q_{pn} \\ \psi &= \begin{bmatrix} 1 & 0 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \Delta t \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{aligned} , \quad (7)$$

  where Δt is the time interval between two epochs and Q_{pn} is the process noise.

- Measurement model F: is used to map a predicted state S^+_{t+1} into a 2D position in image space $M^+_{t+1} = [u, v, d]^T$:

$$M^+_{t+1} = F(S^+_{t+1}) + V_F , \quad (8)$$

  where V_F is the measurement noise.

- Update: let M_{t+1} be the position of a pedestrian which is assigned to τ_j at $(t+1)$. The updated state S_{t+1} and covariance matrix $\Sigma_{SS,t+1}$ are then determined as follows:

$$\begin{aligned} S_{t+1} &= S^+_{t+1} + K(M_{t+1} - F(S^+_{t+1})) \\ \Sigma_{SS,t+1} &= \Sigma^+_{SS,t+1} - K J_F \Sigma^+_{SS,t+1} \\ J_F &= \begin{bmatrix} \frac{-f}{Z} & 0 & \frac{fX}{Z^2} & 0 & 0 \\ 0 & \frac{-f}{Z} & \frac{fY}{Z^2} & 0 & 0 \\ 0 & 0 & \frac{-fb}{Z^2} & 0 & 0 \end{bmatrix} \end{aligned} , \quad (9)$$

  where K is Kalman gain matrix; J_F is Jacobian matrix of F w.r.t the state parameters; and f and b are focal length and baseline of the stereo rig.

Motion model: As people usually smoothly maintain their movements over a short period of time, the velocity of a person can be estimated from a window of k past states:

$$v_{x,t} = \frac{\sum_{l=t-k}^{t} (X_{l+1} - X_l)}{k \Delta t} . \quad (10)$$

The same computation is applied for $v_{z,t}$.

The movement of a pedestrian is usually affected by the behavior of its neighbors. These effects are considered in the motion model to anticipate movements of observed objects in the next epoch. In our model, we define neighbors as persons whose spatial distances are small and moving directions are similar. However, during tracking, some trajectories are not consistent because of wrong assignment or generation from FP detections. Including these incorrect neighbors into the motion model can lead to wrong results. To mitigate this problem, only trajectories with high confidence values are considered as neighbors in our motion model. Let $\{v_i, ..., v_M\}$, $\{\tau_i, ..., \tau_M\}$ be velocities and trajectories of all tracked persons at a certain epoch, respectively. The velocity v_i of each target is predicted as follows:

$$\begin{aligned} v_i &= (\sum_{i=1...M} \omega(\tau_i, \tau_j))^{-1} \sum_{i=1...M} \omega(\tau_i, \tau_j) v_j \\ \omega(\tau_i, \tau_j) &= cos(\varphi_{ij}) e^{-\frac{dis_{ij}}{\epsilon_{dis}}} I_{\varphi_{ij} < \epsilon_\varphi, dis_{ij} < \epsilon_{dis}, \vartheta_{T_j} \geq \epsilon_\vartheta} \end{aligned} , \quad (11)$$

where φ_{ij} and dis_{ij} are angular displacement and spatial distance between two trajectories τ_i and τ_j; I is the indicator

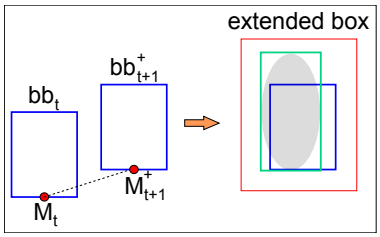

Figure 6. 2D bounding box prediction and correction. The corrected box (green) tightly covers the deactivated target object (gray).

function. ϵ_{φ}, ϵ_{dis}, and ϵ_{ϑ} are threshold parameters, which are used to defined neighbors of τ_j.

3.4 Trajectory extrapolation

During tracking, some pedestrians cannot be detected because of occlusion or visual challenges. Their trajectories are deactivated for a certain amount of epochs. Once a target is deactivated, its position in 3D spaces are still inferred based on its velocity. We then back-project the 3D position into image space to obtain predicted 2D position in the following epochs.

Though this extrapolation helps to recover missing detections, it also increases the number of FP if the continuation of the trajectory is generated from incorrect detections. To address this issue, we estimate the number of epochs ϵ_d that a trajectory can stay in deactivated state before being completely deleted based on the trajectory confidence using the following equation:

$$\epsilon_{d,\tau_{j,t}} = \vartheta^2_{T_{j,t}} a \,, \tag{12}$$

where a is a constant value.

Let S^+_{t+1} and I^+_{t+1} be predicted positions in object and image space of a deactivated trajectory at $(t+1)$. The inferred BB bb^+_{t+1} is determined by moving its previous BB bb_t to a new position such that I^+_{t+1} lies in the middle of the bottom edge of bb^+_{t+1} (see Figure 6). We then check whether the predicted BB contains the tracked pedestrian based on its percentage of pixels that have 3D positions similar to S^+_{t+1}. If most of the 3D points in bb^+_{t+1} lie further away from the camera than the 3D predicted position S^+_{t+1}, we assume that there is no object in bb^+_{t+1}. In the case of a large portion of 3D points nearer to camera than S^+_{t+1}, we assume the object is occluded.

Once the presence of an object in a predicted BB is confirmed, we adjust it by first enlarging the BB and finding all pixels in the extended BB that can belong to that object. The predicted BB is adjusted to cover all those points as shown in Figure 6.

4. EXPERIMENTS AND RESULTS

General goal and dataset: We evaluate the performance of our tracker on the Kitti object tracking benchmark (Geiger et al., 2012). As the ground truth is not provided for the testing data set, we use five different image sequences of the training set to evaluate the effectiveness of sub-components in our framework, namely sequences 13, 15, 16, 17, and 19. To compare the performance of our approach with other state-of-the-art trackers: NOMT (Choi, 2015), RMOT (Yoon et al., 2015), SCEA (Hong Yoon et al., 2016), and CIWT (Ošep et al., 2017). We perform the tracking on the test data set. The evaluation is carried out by the Kitti team.

Parameter	Description	Value
$\epsilon_{\varrho 1}$	detection confidence threshold low	0.25
$\epsilon_{\varrho 2}$	detection confidence threshold high	0.85
ϵ_a	number of active states used to access trajectory confidence	20
ϵ_α	angular displacement threshold	$\frac{\pi}{3}$
ϵ_{dis}	distance threshold of two neighbors	2m
ϵ_ϑ	trajectory confidence threshold	0.8
A	deactivate states constance	10
α, β, γ	weight parameters in Equation (6)	0.5, 0.2, 0.3
ρ, θ, ν, ι	weight parameters in Equation (3)	0.1, 0.6, 0.1, 0.2

Table 1. Setting of parameters of our tracking system.

Detections	Recall ↑	FP ↓	Precision ↑
mask R-CNN low DCT $\epsilon_\varrho = 0.25$	67.94	51.79	56.74
mask R-CNN high DCT $\epsilon_\varrho = 0.85$	64.06	**15.06**	**80.96**
Ours $\epsilon_{\varrho 1} = 0.25$, $\epsilon_{\varrho 2} = 0.85$	**75.74**	19.6	79.44

Table 2. The comparison of detection results.

Evaluation metrics: The performance of our tracker is analyzed using the CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008). The tracking accuracy MOTA is computed from three types of errors: false negative (FN), FP, and Id switch (IDs). The localization error MOTP is measured by the intersection over union between tracked objects and ground truth bounding boxes in image space. We compute 3D-MOTP to assess the estimated positions of tracked pedestrians in 3D object space as well. In addition, we also utilize four additional metrics including the percentage of most tracked (MT) and most lost (ML) trajectories, the number of Id switches (IDs) and fragmentation (FR) to compare our method against the state-of-the-art (Li et al., 2009).

Parameters setting: The thresholds and weight parameters used in our equations are determined heuristically and applied for all image sequences. Their values are listed in Table 1.

Detection results: Unlike the conventional approach, which uses only one fixed threshold for selecting TP detections, in our TCD method, we use both low ($\epsilon_{\varrho 1} = 0.25$) and high ($\epsilon_{\varrho 2} = 0.85$) DCTs. Table 2 shows the comparison of our detection results with those of two single-threshold mask R-CNN computations. It is evident that even with a very low DCT $\epsilon_\varrho = 0.25$, the mask R-CNN just obtains 67.94 % recall, while our approach which combine both the TCD and trajectory extrapolation methods can improve it to 75.74 %. Moreover, while increasing the recall value, our tracker also keeps the number of FPs at a comparably low rate of 19.6 %, resulting in a high precision of 79.44, which is very similar to the value of mask R-CNN with a very high DCT $\epsilon_\varrho = 0.85$, and much better than mark R-CNN with a low DCT.

Moreover, the MOTA value may actually be even higher because a number of the Kitti reference bounding boxes are not very accurately placed: they do not cover the appearance of pedestrians in the images as tightly as our tracker (see Figure 7). Therefore, in order to get a better insight on the performance of our method, we evaluate the results of our tracker in 3D space with difference intersection over union thresholds as illustrated in Figure 8. With a IoU decreased from 0.5 to 0.4, the MOTA and MT values are improved significantly, while 3D-MOTP with one meter threshold stays nearly constant. Therefore,

Figure 7. Examples of inaccurate references of the Kitti data (green boxes). Our detections (red boxes) cover the pedestrians better.

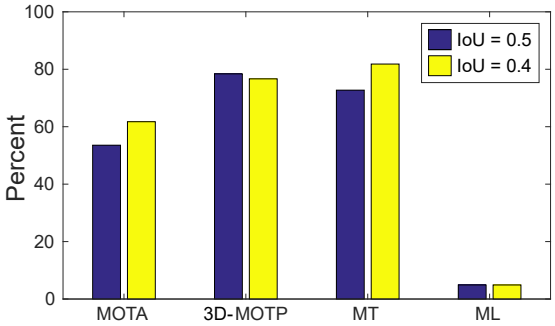

Figure 8. Performance of our tracker in 3D object space with different IoU thresholds.

in our tracking system, pedestrian trajectories can be well estimated even with a lower IoU than 0.5 as defined by the Kitti benchmark. This is because either some of references are not precise or 3D accurate localization does not require a very accurate detection, which needs to cover the entire interesting object.

Furthermore, Figure 9 shows that the localization in 3D object space achieves best results when the distances between tracked objects and camera fall in the range 5–15 m. This is also the critical distance for vehicles to stop when reactions are required.

Tracking results: The tracking performance of our approach (CAT) and other state-of-the-art methods on the Kitti test data set are presented in Table 3[1]. It can be observed that our method shows comparable results to NOMT in most of the metrics, except for the number of ID switch. This is probably because NOMT uses additional optical flow features in the data association step and solves the assignment problem with a temporal window, while we just employ appearance and geometry properties to link detections in two consecutive frames. Compared to CIWT, RMOT, and SCEA, our method demonstrates remarkable improvements in MOTA, MT, and ML. This is mainly because we used different detection methods. In addition, our TCD and extrapolation with confidence awareness approaches enable the achievement of high recall together with high precision of detection, which strongly affects the MOTA value. In Table 3, it is obvious that SCEA has a MT number larger than ours, which means that the number of different pedestrians are tracked by our tracker is much higher. This is an important reason why they can achieve such a low number of ID switch and FR compare to us.

5. CONCLUSION

Pedestrian tracking still remains a highly challenging problem, mainly due to noisy detection results and crowded scenes

[1] www.cvlibs.net/datasets/kitti/eval_tracking.php

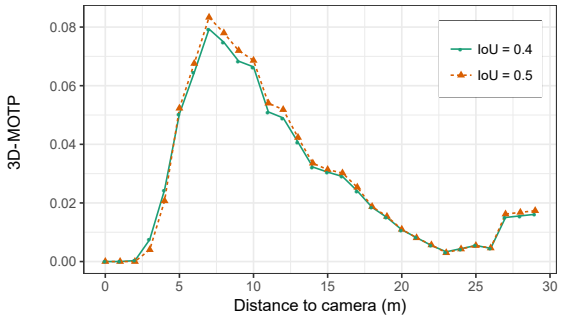

Figure 9. The histogram of 3D-MOTP w.r.t the distance between tracked pedestrians and the stereo rig.

Tracker	MOTA ↑	MOTP ↑	MT ↑	ML ↓	IDs ↓	FR ↓
CAT (ours)	52.35	71.57	**34.36**	23.71	206	804
NOMT	**57.67**	**72.17**	**34.36**	**19.24**	108	799
RMOT	43.77	71.02	19.59	41.24	153	748
SCEA	43.91	71.86	16.15	43.30	**56**	**641**
CIWT	43.37	71.14	13.75	34.71	112	901

Table 3. Evaluation results on the Kitti data set of our tracking method CAT and other state-of-the-art methods.

with many occlusions. With the goal of improving the tracking results, we proposed a framework to track pedestrian in 3D object space with the awareness of both, detection and trajectory confidence values. Moreover, employing the power of existing CNNs in different stages of our tracker is also an important factor to achieve better tracking performance.

The evaluation results on the Kitti dataset demonstrate that our tracking approach is at least comparable to the state-of-the-art methods. We can obtain both high recall and precision results, which leads to a noticeable increase of MOTA (52.35 %), MT (30.36 %), and ML (23.71 %). Additionally, our tracker can also localize pedestrians in 3D object space precisely (within 1 meter) even in cases where just a portion of the target object can be observed.

In future work we will further investigate the values of weight parameters in Equation (5) and Equation (6) to evaluate their influence on the performance of our tracker. In addition, the relations of all tracked objects are more or less maintained during consecutive epochs. Therefore, tracking pedestrians with the consideration of neighbours as constraints when solving the association and localization problems can help to improve the tracking accuracy. Finally, our framework can be extended to track pedestrians and other objects from multiple viewpoints in the context of collaborative autonomous cars.

ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG) as part of the Research Training Group i.c.sens [RTG 2159].

REFERENCES

Berclaz, J., Fleuret, F., Turetken, E. and Fua, P., 2011. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(9), pp. 1806–1819.

Bernardin, K. and Stiefelhagen, R., 2008. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*.

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E. and Van Gool, L., 2011. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(9), pp. 1820–1833.

Choi, W., 2015. Near-online multi-target tracking with aggregated local flow descriptor. In: *IEEE international conference on computer vision (ICCV 2015)*, pp. 3029–3037.

Dehghan, A., Modiri Assari, S. and Shah, M., 2015. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pp. 4091–4099.

Fagot-Bouquet, L., Audigier, R., Dhome, Y. and Lerasle, F., 2016. Improving multi-frame data association with sparse representations for robust near-online multi-object tracking. In: *European Conference on Computer Vision (ECCV 2016)*, Springer, pp. 774–790.

Geiger, A., Lenz, P. and Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, pp. 3354–3361.

Gelb, A., 1974. *Applied Optimal Estimation*. MIT press.

He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn. In: *IEEE International Conference on Computer Vision (ICCV 2017)*, pp. 2980–2988.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 770–778.

Henschel, R., Leal-Taixé, L., Cremers, D. and Rosenhahn, B., 2018. Fusion of head and full-body detectors for multi-object tracking. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1540–1550.

Hermans, A., Beyer, L. and Leibe, B., 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hong Yoon, J., Lee, C.-R., Yang, M.-H. and Yoon, K.-J., 2016. Online multi-object tracking via structural constraint event aggregation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 1392–1400.

Jafari, O. H., Mitzel, D. and Leibe, B., 2014. Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras. In: *IEEE International Conference on Robotics and Automation (ICRA 2014)*, pp. 5636–5643.

Kieritz, H., Becker, S., Hübner, W. and Arens, M., 2016. Online multi-person tracking using integral channel features. In: *13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2016)*, pp. 122–130.

Klinger, T., Rottensteiner, F. and Heipke, C., 2017. Probabilistic multi-person localisation and tracking in image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing* 127, pp. 73–88.

Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B. and Savarese, S., 2014. Learning an image-based motion context for multiple people tracking. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, pp. 3542–3549.

Leal-Taixé, L., Milan, A., Schindler, K., Cremers, D., Reid, I. and Roth, S., 2017. Tracking the trackers: An analysis of the state of the art in multiple object tracking. *arXiv preprint arXiv:1704.02781*.

Lenz, P., Geiger, A. and Urtasun, R., 2015. Followme: Efficient online min-cost flow tracking with bounded memory and computation. In: *IEEE International Conference on Computer Vision (ICCV 2015)*, pp. 4364–4372.

Li, Y., Huang, C. and Nevatia, R., 2009. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 2953–2960.

Linder, T., Breuers, S., Leibe, B. and 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 2016)*, pp. 5512–5519.

Mitzel, D., Horbert, E., Ess, A. and Leibe, B., 2010. Multi-person tracking with sparse detection and continuous segmentation. In: *European Conference on Computer Vision (ECCV 2010)*, pp. 397–410.

Nguyen, U., Rotteinsteiner, F. and Heipke, C., 2018. Object proposals for pedestrian detection in stereo images. *38. Wissenschaftlich-Technische Jahrestagung der DGPF und PFGK18 Tagung in München* Band 27(9), pp. 611–623.

Ošep, A., Mehner, W., Mathias, M. and Leibe, B., 2017. Combined image- and world-space tracking in traffic scenes. In: *IEEE International Conference on Robotics and Automation (ICRA 2017)*, pp. 1988–1995.

Pirsiavash, H., Ramanan, D. and Fowlkes, C., 2011. Globally-optimal greedy algorithms for tracking a variable number of objects. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pp. 1201–1208.

Schindler, K., Ess, A., Leibe, B. and Van Gool, L., 2010. Automatic detection and tracking of pedestrians from a moving stereo rig. *ISPRS Journal of Photogrammetry and Remote Sensing* 65(6), pp. 523–537.

Xiang, Y., Alahi, A. and Savarese, S., 2015. Learning to track: Online multi-object tracking by decision making. In: *IEEE International Conference on Computer Vision (ICCV 2015)*, pp. 4705–4713.

Yamaguchi, K., McAllester, D. and Urtasun, R., 2014. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In: *European Conference on Computer Vision (ECCV 2014)*, Springer, pp. 756–771.

Yoon, J. H., Yang, M.-H., Lim, J. and Yoon, K.-J., 2015. Bayesian multi-object tracking using motion context from multiple objects. In: *IEEE Winter Conference on Applications of Computer Vision (WACV 2015)*, pp. 33–40.

Zamir, A. R., Dehghan, A. and Shah, M., 2012. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In: *European Conference on Computer Vision (ECCV 2012)*, Springer, pp. 343–356.

Zhang, L. and van der Maaten, L., 2013. Structure preserving object tracking. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, pp. 1838–1845.

Zhang, L., Li, Y. and Nevatia, R., 2008. Global data association for multi-object tracking using network flows. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8.

Zhang, S., Benenson, R., Omran, M., Hosang, J. and Schiele, B., 2016. How far are we from solving pedestrian detection? In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 1259–1267.