People Tracking in RGB-D Data With On-line Boosted Target Models

Matthias Luber Luciano Spinello Kai O. Arras

***Abstract*— People tracking is a key component for robots that are deployed in populated environments. Previous works have used cameras and 2D and 3D range finders for this task. In this paper, we present a 3D people detection and tracking approach using RGB-D data. We combine a novel multi-cue person detector for RGB-D data with an on-line detector that learns individual target models. The two detectors are integrated into a decisional framework with a multi-hypothesis tracker that controls on-line learning through a track interpretation feedback. For on-line learning, we take a boosting approach using three types of RGB-D features and a confidence maximization search in 3D space. The approach is general in that it neither relies on background learning nor a ground plane assumption. For the evaluation, we collect data in a populated indoor environment using a setup of three Microsoft Kinect sensors with a joint field of view. The results demonstrate reliable 3D tracking of people in RGB-D data and show how the framework is able to avoid drift of the on-line detector and increase the overall tracking performance.**

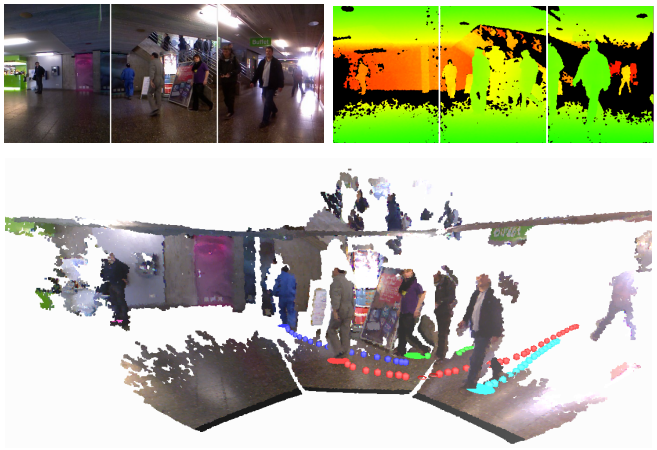

Fig. 1. People tracking in RGB-D data. The top pictures show the three color and depth images, below the 3D point cloud. The data was collected in the lobby of a large university canteen at lunch time with a setup joining the views of three Kinect sensors. The colored disks and dots in the point cloud show the positions and trajectories of five tracked persons.

I. Introduction

People detection and tracking is an important and fundamental component for many robots, interactive systems and intelligent vehicles. Popular sensors for this task are cameras and range finders. While both sensing modalities have advantages and drawbacks, their distinction may become obsolete with the availability of affordable and increasingly reliable RGB-D sensors that provide both image and range data.

Many researchers in robotics have addressed the issue of detection and tracking people in range data. Early works were based on 2D data in which people have been detected using ad-hoc classifiers that find moving local minima in the scan [1], [2]. A learning approach has been taken by Arras *et al.* [3], where a classifier for 2D point clouds has been trained by boosting a set of geometric and statistical features.

People detection and tracking in 3D range data is a rather new problem with little related work. Navarro *et al.* [4] collapse the 3D scan into a virtual 2D slice to find salient vertical objects above ground and classify a person by a set of SVM classified features. Bajracharya *et al.* [5] detect people in point clouds from stereo vision by processing vertical objects and considering a set of geometrical and statistical features of the cloud based on a fixed pedestrian model. Unlike these works that require a ground plane assumption, Spinello *et al.* [6] overcome this limitation via a voting approach of classified parts and a top-down verification procedure that learns an optimal feature set and volume tessellation.

In the computer vision literature, the problem of detecting, tracking and modeling humans has been extensively studied [7], [8], [9], [10]. A major difference to range-based systems is that the richness of image data makes is straightforward to learn target appearance models. For this reason, visual tracking systems can achieve good results with methods as simple as independent particle filters with nearest-neighbor data association [11]. Dense depth data from stereo are used by Beymer and Konolige [12] to support foreground segmentation in an otherwise vision-based people detection and tracking system. They use a set of binary person templates to detect people in images and demonstrate multi-person tracking with learned appearance-based target models. The work of [13], [14] detect people in intensity images and track them in 3D. In [15] a stereo system for combining intensity images, stereo disparity maps, and optical flow is used to detect people. Multi modal detection and tracking of people is performed in [16] where a trainable 2D range data and camera system is presented.

This paper advances the state-of-the-art in the following aspects. First, we address the problem of detecting

All authors are with the Social Robotics Lab, Department of Computer Science, University of Freiburg, Germany {luber,spinello,arras}@informatik.uni-freiburg.de.

and tracking people in RGB-D data. We combine a generic person detector with an on-line learned person detector and a multi-hypothesis tracker (MHT), able to estimate the motion state of multiple people in 3D. On-line learning is a recent aspect from the perspective of range data-based object tracking that usually deals with targets of identical appearance. We present a novel framework to integrate the two detectors and the tracker that is able to bridge gaps of misdetections of the generic detector and handle occlusions while avoiding drift of the on-line detector by a track interpretation feedback from the MHT. We give quantitative results using the CLEAR MOT performance metric. Then, we adapt the on-line learning method for target appearances by Grabner *et al.* [17] to RGB-D data and present a decisional framework for its integration into the MHT in which information from the tracker is fed back to control the on-line learner and combine it with the a priori person detector. Unlike all the previous work we consider image and range data as two equally important cues for detection, tracking, and target model adaptation, all soundly integrated into a single MHT framework.

The paper is structured as follows: the generic people detector is briefly summarized in the next section followed by the description of our on-line AdaBoost learning approach for target appearances in RGB-D data in Section III. The integration of this learning procedure into the tracking system is described in Section IV. Section V describes the experiments and gives the results. Section VI concludes the paper.

II. Detection of People in 3D Range Data

In this section we briefly summarize the generic people detector used in this paper. We rely on a novel RGB-D person detector called Combo-HOD (Combined Histograms of Oriented Depths and Gradients). The method takes inspiration from Histogram of Oriented Gradients (HOG) introduced by Dalal and Triggs [7] and combines the HOG detector in the color image with a novel approach in the depth image called Histograms of Oriented Depths (HOD).

Since RGB-D data contains both color and depth information, the Combo-HOD detector combines the two sensory cues. HOD descriptors are computed in the depth image and HOG descriptors are computed in the color image. They are fused on the level of detections via a weighted mean of the probabilities obtained by a sigmoid fitted to the SVM outputs. HOD includes a depth-informed scale-space search in which the used scales in an image are first collected and then tested for compatibility with the respective depth. This test is made particularly efficient by the use of integral tensors, an extension of integral images over several scales. This strategy dramatically reduces the number of descriptors computed in the image at improved detection rates. For more details, the reader is referred to [18].

The output of the detector in each step are the positions and size of all targets in 3D space and the center and size of the bounding boxes in the depth images. They are the observations $z_i(t)$ that constitute the set of m_k observations $\mathcal{Z}(t)$ at time index t.

III. On-line Boosting

The detector described in the previous section learns a generic person model from a priori labeled data. In this section, we describe the use of on-line boosting to learn target appearance models in RGB-D data, later used to guide data association in the tracking system.

Boosting is a widely used technique to improve the accuracy of learning algorithms. Given training samples $\mathbf{x}$ with labels y, a strong classifier $H(\mathbf{x})$ is computed as linear combination of a set of weighted hypotheses called weak classifiers $h(\mathbf{x})$. The discrete AdaBoost algorithm by Freund and Shapire [19] belongs to the most popular boosting algorithms. The method trains weak classifiers from labeled training samples $(\mathbf{x}, y)$, initialized with uniform weights w_i associated to each $\mathbf{x}$. Learning is done in rounds where the weights are updated based on the mistakes of the previous weak learner. By increasing the weights of the wrongly classified samples the algorithm focuses on the difficult examples.

On-line boosting, initially proposed by Oza and Russell [20], processes each training instance "on arrival" without the need of storage and reprocessing, and maintains a current hypothesis that reflect all the training samples seen so far. The approach has been applied for object detection while tracking by Grabner *et al.* [17]. We build upon the latter to develop our on-line people detector in RGB-D data.

A. Updating the Weak Classifiers

Unlike the off-line approach to boosting, the on-line algorithm presents training samples only once and discards them after training. The weak classifiers have thus to be updated in an on-line fashion each time a new training sample is available. As the difficulty of the samples is not known in advance the computation of the weight distribution of the samples is a critical issue. The basic idea of on-line boosting is that the weight of a sample (called importance λ in this context) can be estimated by propagating it through a fixed chain of weak classifiers [20]. If the sample is misclassified, λ is increased proportional to the error of the weak classifier. Therefore, the importance has the same effect as the adapted weight in the off-line approach. The error of the i-th weak classifiers is estimated from the summed weights of the correctly (λ_i^{corr}) and wrongly (λ_i^{wrong}) classified samples,

$$e_i = \frac{\lambda_i^{wrong}}{(\lambda_i^{wrong} + \lambda_i^{corr})}. \quad (1)$$

B. On-line-boosting for Feature Selection

For the purpose of learning target models during tracking, Grabner *et al.* [17] propose *feature selectors*. The main idea is to apply on-line boosting not directly to

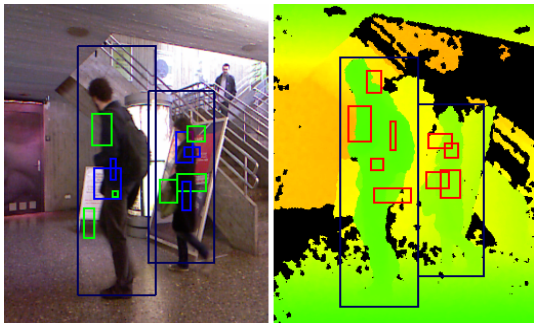

Fig. 2. Bounding boxes of two detected persons in the RGB and depth images. The ten best features of each on-line detector are marked with colored rectangles. Haar-like features calculated on the intensity image are shown in green and Haar-like features computed on the depth image are marked in red. The *Lab* color features calculated on the RGB image are depicted in blue.

the weak classifiers but to the selectors. A selector h^{sel} selects the best weak classifier from a pool of M weak learners $\mathcal{F}$ with 'best' being defined by the lowest error.

With the number of selectors N being a fix parameter, the following procedure is repeated for all selectors when a new sample $(\mathbf{x}, y)$ arrives: First, all weak classifiers are updated and the best one, denoted m^+, is selected

$$h_n^{sel}(\mathbf{x}) = h_{m^+}^{weak}(\mathbf{x}) \tag{2}$$

with $m^+ = \arg\min_m(e_{n,m})$ and $e_{n,m}$ defined like Eq. 1 with subscript n, m for i. Then, the voting weight $\alpha_n = \frac{1}{2} \cdot \ln(\frac{1-e_n}{e_n})$ is computed where $e_n = e_{n,m^+}$ and the updated importance weight λ is propagated to the next selector h_{n+1}^{sel}. Similar to AdaBoost, λ is increased if h_n^{sel} predicts $\mathbf{x}$ correctly and decreased otherwise.

The strong classifier is finally obtained by computing the confidence as a linear combination of the N selectors and applying the signum function,

$$\kappa(\mathbf{x}) = \sum_{n=1}^{N} (\alpha_n \cdot h_n^{sel}(\mathbf{x})), \qquad H(\mathbf{x}) = sign(\kappa(\mathbf{x})). \tag{3}$$

Unlike the off-line version, the on-line procedure creates an always-available strong classifier in a any-time fashion.

In order to increase the diversity of the classifier pool $\mathcal{F}$ and to adapt to appearance changes of the targets, at the end of each iteration, the worst weak classifier is replaced by one randomly chosen from $\mathcal{F}$.

C. Features

We take advantage of the richness of RGB-D data by computing three types of features that correspond to the weak classifiers: Haar-like features [21] in the intensity image (converted from the RGB values), Haar-like features in the depth image, and illumination agnostic *Lab* color features in the RGB image. *Lab* features are computed by summing up the intensity values in a* (b*) space under the area. The advantage of the *Lab* color model is that features in a* or b* space can compactly and robustly subsume entire RGB histograms. A total of M features is computed where the initial number of features is $M/3$ for all types. Given the above mentioned adaptation mechanism, their relative numbers can change to best describe a target dynamically.

The features are computed in rectangular areas sampled with randomized positions and scales in the bounding box associated to each target. This is done once at initialization and then kept fix over the lifetime of a target (up to the weak feature that get replaced). The best ten features of two persons are shown in Fig. 2.

D. On-line Boosting for Tracking

On-line boosting enables a tracker to continuously update a target model to optimally discriminate it from the current background. This is a formulation of tracking as a classification problem [22] which is implemented by a confidence maximization procedure around the current tracking region . The region is obtained as the bounding box of the previous detection. All features within the region are considered the positively labeled foreground samples. The negative samples are obtained by sweeping the bounding box over a local neighborhood. The classifier is then evaluated at each sweep position of this neighborhood yielding a confidence map whose maximum is taken as the new position of the tracking region. The classifier is updated in this region and the process is continued. The evolution of the confidence values over time can be seen in Fig. 5.

Unlike [17] where the new region is bootstrapped from the previous detection, we use the bounding box position of the a priori detector to recenter the on-line detector. This strategy avoids a key problem of on-line adaptation namely drifting of the model to background, clutter, or other targets.

IV. Integration into the Tracking System

In this section we describe how the on-line detector is integrated into a Kalman filter based multi-hypothesis tracking framework (MHT). For reasons of limited space, we will only discuss the aspects that change in the MHT, refer to [23], [24] for more details.

In short, the MHT algorithm hypothesizes about the target states by considering all statistically feasible assignments between measurements and tracks and all possible interpretations of measurements as false alarms or new track and tracks as matched, occluded or obsolete. Thereby, the MHT handles the entire life-cycle of tracks from creation and confirmation to occlusion and deletion.

Formally, let $\xi(t) = (x_t\ y_t\ z_t\ \dot{x}_t\ \dot{y}_t\ \dot{z}_t)^T$ be the filtered state of a track $\mathbf{t}$ at time t with position and velocity information in 3D and Σ its associated 6×6 covariance. Let $Z(t) = \{\mathbf{z}_i(t)\}_{i=1}^{m_t}$ be the set of m_t observations which in our case is the set of detected people in RGB-D data. Observations consist in a 3D position from the a priori detector $z_i(t)$ and a training sample $\mathbf{x}_i(t)$ from the on-line detector. The sample $\mathbf{x}_i(t))$ is a vector of stacked features values computed in the rectangular areas within the current tracking region.

Let $\Omega_l(t)$ be the l-th hypothesis at time t and $\Omega_{p(l)}^{t-1}$ the parent hypothesis from which $\Omega_l(t)$ was derived. Let

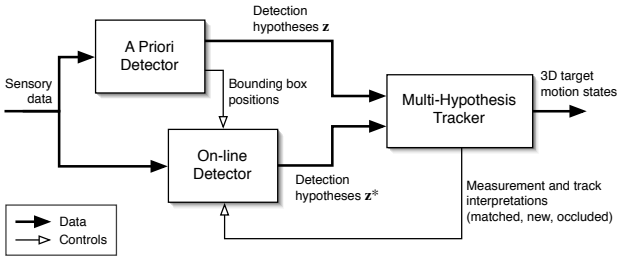

Fig. 3. The decisional framework to integrate both detectors and the tracking system

Fig. 4. The setup consisting in three vertically mounted Kinect sensors offering a joint field of view of $130° \times 50°$ and supplying RGB-D data with a resolution of 1440×640 pixels at $30\,Hz$. They are mounted at $1.2\,m$ height.

further $\psi_j(t)$ denote a set of assignments which associates predicted tracks to measurements in $Z(t)$. In each cycle, the method tries to associate the tracks in the parent hypotheses of the previous step to the new set $Z(t)$, producing all possible assignment sets $\psi(t)$ that each give birth to a child hypothesis that branches off its parent. This results in an exponentially growing hypothesis tree. Most practical MHT implementations prune the tree by Murty's algorithm able to generate and evaluate the current k best hypotheses in polynomial time.

A. Joint Likelihood Data Association

The measurement likelihood in the regular MHT $p(z_i(t)|\psi_j^t, \Omega_{p(l)}^{t-1})$ consists in two terms, one for observations interpreted as new tracks and false alarms (which we leave unchanged) and a second one for matched observations $z_i(t)$ that follows the Gaussian likelihood model centered on the measurement prediction $\hat{z}_j(t)$ with innovation covariance matrix $S_{ij}(t)$, $p(z_i(t)|\psi_j^t, \Omega_{p(l)}^{t-1}) = \mathcal{N}(z_i(t)\,; \hat{z}_j(t), S_{ij}(t))$. This likelihood quantifies how well an observation matches a predicted measurement based on position and velocity.

Here, the on-line classifier H adds an appearance likelihood that expresses how much the observed target's appearance matches the learned model. We thus have a joint likelihood that accounts for both motion state and appearance. With $\mathbf{x}_i(t)$ being the feature descriptor of $z_i(t)$, $\mathbf{z}_i(t) = (z_i(t), \mathbf{x}_i(t))$, and assuming independence between the two terms,

$$p(\mathbf{z}_i(t)|\psi_j^t, \Omega_{p(l)}^{t-1}, H^{t-1}) = p(z_i(t)|\psi_j^t, \Omega_{p(l)}^{t-1}) \cdot p(\mathbf{x}_i(t)|H^{t-1}). \quad (4)$$

We also model the appearance likelihood to be a Gaussian pdf centered on the maximum confidence of the strong classifier (which is 1)

$$p(\mathbf{x}_i(t) \mid H^{t-1}) = \mathcal{N}(\kappa(\mathbf{x}_i(t))\,; 1, \sigma_a^2), \quad (5)$$

where σ_a^2 is the variance of the Gaussian and a smoothing parameter to trade off the two likelihoods.

B. Feeding Data Association Back to On-line Boosting

In each cycle, the tracker produces assignments of measurements to tracks and interpretations of measurements as new tracks or false alarms and of track as occluded or deleted. This information directly serves the on-line boosting algorithm to create and update the strong classifiers:

- When an observation $\mathbf{z}_{new}$ is declared as a new target, a new track $\mathbf{t}_{new}$ is initialized and a new strong classifier H_{new} is created at the bounding box position of the hypothesis of the a priori detector.
- When an existing target $\mathbf{t}_i$ is associated to an observation $\mathbf{z}_j(t)$, the strong classifier H_i^t is updated using the features $\mathbf{x}_j(t)$ calculated within the new bounding box of the a priori detector. The on-line detector is centered at this new bounding box position.
- When the MHT declares a track as occluded, there are two possible reasons: an occlusion or a misdetection. To cope with both cases, we proceed as follows: Given the on-line learned model, we search for targets without valid observations by centering a 3D confidence map around the motion prediction of the Kalman filter. The map size is proportional to the uncertainty of the prediction, the confidence values are calculated using the projections of the 3D positions into image space. If a high-confidence match can be found, we interpret the event as a misdetection and make the confidence maximum an observation $\mathbf{z}^*(t)$. Otherwise, we interpret the event as a target occlusion and stop on-line learning of the corresponding strong classifier until the target reappears. This strategy also avoids drifting of the model to background, clutter, or other targets.

Observations $\mathbf{z}^*(t)$ from the on-line detector are treated like regular observations for the MHT.

V. Experiments

To evaluate and compare the different detector approaches, we collected a large-scale indoor data set with unscripted behavior of people. The data set has been taken in the lobby of a large university canteen at lunch time. The a priori detector has been trained with an additional background data set collected in another, visually different university building. This is to avoid detector bias towards the visual appearance of the canteen lobby, especially since we acquired the data from a stationary sensor. The data set has been manually annotated to include the bounding box in 2D depth image space, the

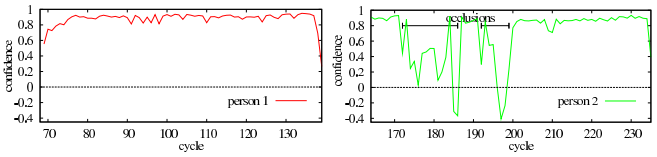

Fig. 5. Evolution of the confidence of the on-line detector. The top image shows the confidences over the life cycle of a track. After initialization the values achieves steady state. Person 2 is occluded twice between frames 172 to 185 and frames 192 to 199. Thanks to the feedback from the MHT tracker, the on-line detector pauses its adaptation. This strategy avoids drifting of the model to background, clutter, or other targets. When the person reappears, adaptation is resumed immediately with high confidence.

visibility of subjects (fully visible/partially occluded), and the data association ground truth of the tracks. A total of 3021 instances of people in 1133 frames and 31 tracks have been labeled. The data set will be made available on the laboratory webpage at publication date of this paper.

The sensory setup for data collection is shown in Fig. 4. It consists in three vertically mounted Kinect sensors that jointly extend the field of view to $130° \times 50°$. Measures have been taken to calibrate the intrinsics and extrinsics of the setup and to guarantee synchronized acquisition of the three images at frame rate.

The parameters of the MHT have been learned from a training data set over 600 frames. The detection probability is set to $p_{det} = 0.99$ and the termination likelihood to $\lambda_{del} = 30$. The average rates of new tracks and false alarms are determined to be $\lambda_{new} = 0.001$ and $\lambda_{fal} = 0.005$, respectively. Further, the maximal number of hypothesis N_{Hyp} is set to 100. The strong classifiers of the targets are based on 50 selectors which are trained with 50 weak hypotheses.

To assess the impact of the on-line boosting onto the tracking performance we run the tracker with the a priori detector only to obtain a baseline. All following runs are then compared using the CLEAR MOT metrics [25]. The metric counts three numbers with respect to the ground truth that are incremented at each frame: misses (missing tracks that should exist at a ground truth position, FN), false positives (tracks that should not exist, FP), and mismatches (track identifier switches, ID). The latter value quantifies the ability to deal with occlusion events that typically occur when tracking people. From these numbers, two values are determined: MOTP (avg. metric distance between estimated targets and ground truth) and MOTA (avg. number of times of a correct tracking output with respect to the ground truth). We ignore MOTP as it is based on a metric ground truth of target positions which is unreliable in our data.

A. Results

First, we analyze the confidence values of the strong classifier H and the integration framework in different situations (see Fig. 5). Person 1 traverses the sensor field of view without interference with other targets. After an initialization phase of nearly ten frames, the on-line

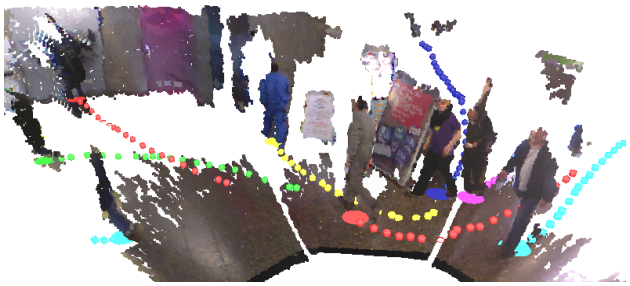

Fig. 6. Visualization of the 3D point cloud produced by the three Kinect sensors including the positions and trajectories of eight of 31 tracks in the data set. The colored disks mark the current Kalman filter estimates of the target positions, the small dots show their past trajectories. The tracker maintains full 3D estimates as it can be seen by the dark blue trajectory of the subject coming down the stairs.

detector has adapted to its appearance and achieves steady state at a value of around 0.8. Person 2 undergoes two occlusions. During the occlusions the confidence values drop immediately, indicating that the target is no longer visible. As the MHT correctly declares the target as occluded, adaptation of H is paused and resumed with high confidences after the person reappears. We have further investigated the usage statistics of the three features types of the on-line detector. They are generally used with similar frequency and importance.

We then compare the on-line boosting approach to the baseline using the CLEAR MOT metrics. The results show a clear improvement of all values except for the number of false positives (see Table I). We manually inspected the behavior of the tracker and discuss the insights gained.

The strongest impact of the presented approach is the reduction of the number of missed targets by 50%. This improvement is caused by the on-line observations $\mathbf{z}^*$. When the a priori detector fails to detect an existing track in several consecutive frames, the best MHT hypothesis will eventually (and wrongly) declare the track as deleted. When this happens, the miss count (FN) is increased at each frame until the detector finds the target again and creates a new track. This is where the $\mathbf{z}^*$ observations come into play by detecting the target from the on-line learned model. Given a $\mathbf{z}^*$, the MHT can match the target and correctly continue the track.

This benefit comes at the expense of a delayed deletion of tracks that are incorrectly created from wrong false positives of the a priori detector. In this case, the on-line detector tries to continue the track with the same strategy leading to a increase of the number of false positives (FP) by 19%. Due to a limited RGB-D training set the a priori detector tends to detect false positives at a certain cluttered location in space. We manually removed these detections that bias the tracking performance metric by 683 FP for the baseline and 949 FP for the on-line boosting approach.

The improvement in the number of id switches (ID) is achieved by the joint likelihood model that guides data association in situations of interacting and thus occluding

	FN	FP	ID	MOTA
Baseline	1502	168	42	62%
On-line boosting	751	201	32	78%
Improvement	50%	-19%	24%	16%

TABLE I
CLEAR MOT RESULTS.

targets. The fact that this number is not higher is due to the unscripted behavior of people in our data set. At the particular place of data collection, subjects mainly walked past rather than creating situations that stress the occlusion handling capability of the tracker.

VI. Conclusions

In this paper we presented a novel 3D people detection and tracking approach in RGB-D data. We combined on-line learning of target appearance models using three types of RGB-D features with multi-hypothesis tracking. We propose an decisional framework to integrate the on-line detector, an a priori off-line learned detector of people and a multi-hypothesis tracker. The framework enables the tracker to support the on-line classifier in training only on the correct samples and to guide data association through a joint motion and appearance likelihood. It also avoids the key problem of on-line adaptation namely drifting of models to background, clutter, or other targets by resetting the detection window at the location of the a priori detector and pausing adaptation in case of occlusions. The framework further allows to fill gaps of false negatives from the a priori detector by observations of the on-line detector.

The experiments show a clear overall improvement of the tracking performance, particularly in the number of missed tracks and also in the number of identifier switches. They demonstrate that the on-line classifier contributes to find the correct observations in cases when the a priori detector fails. This reduces the number of missed tracks by 50%. Further, the joint data association likelihood decreases the number of track identifier switches by 24%. The overall tracking accuracy (MOTA) is improved by 16%.

Future work will focus on the collection and annotation of more RGB-D data sets containing a variety of social situations that stress more aspects of this approach. Finally, the target models are currently learned for each track in isolation. We plan to extend the on-line detector to learn the models jointly over all tracks to even better distinguish them from each other.

Acknowledgment

This work has been supported by the German Research Foundation (DFG) under contract number SFB/TR-8.

References

[1] A. Fod, A. Howard, and M. Mataríc, "Laser-based people tracking," in *Int. Conf. on Robotics and Automation (ICRA)*, 2002.

[2] D. Schulz, W. Burgard, D. Fox, and A. Cremers, "People tracking with a mobile robot using sample-based joint probabilistic data association filters," *International Journal of Robotics Research (IJRR)*, vol. 22, no. 2, pp. 99–116, 2003.

[3] K. O. Arras, O. Martínez Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *Int. Conf. on Robotics and Automation (ICRA)*, 2007.

[4] L. Navarro-Serment, C. Mertz, and M. Hebert, "Pedestrian detection and tracking using three-dimensional LADAR data," in *International Conference on Field and Service Robotics*, Cambridge, USA, 2009.

[5] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L. Matthies, "Results from a real-time stereo-based pedestrian detection system on a moving vehicle," in *Workshop on People Detection and Tracking, IEEE ICRA*, Kobe, Japan, 2009.

[6] L. Spinello, M. Luber, and K. O. Arras, "Tracking people in 3D using a bottom-up top-down people detector," in *Int. Conf. on Robotics and Automation (ICRA)*, Shanghai, China, 2011.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, San Diego, USA, 2005.

[8] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, San Diego, USA, 2005.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, Anchorage, USA, 2008.

[10] M. Enzweiler and D. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. on Pattern Analysis and Machine Intell. (PAMI)*, vol. 31, no. 12, 2009.

[11] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multi-person tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. on Pattern Analysis and Machine Intell. (PAMI)*, vol. 33, no. 9, 2011.

[12] D. Beymer and K. Konolige, "Real-time tracking of multiple people using stereo," in *ICCV Workshop on Frame-rate Applications*, Kerkyra, Greece, 1999.

[13] A. Ess, B. Leibe, K. Schindler, and L. V. Gool, "Robust multiperson tracking from a mobile platform," *IEEE Trans. on Pattern Analysis and Machine Intell. (PAMI)*, vol. 31, no. 10, pp. 1831–1846, 2009.

[14] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool, "Coupled object detection and tracking from static cameras and moving vehicles," *IEEE Trans. on Pattern Analysis and Machine Intell. (PAMI)*, pp. 1683–1698, 2008.

[15] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, 2010.

[16] L. Spinello, R. Triebel, and R. Siegwart, "Multiclass multimodal detection and tracking in urban environments," *Int. Journal of Robotics Research*, vol. 29, no. 12, pp. 1498–1515.

[17] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, New York, USA, 2006.

[18] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, San Francisco, USA, 2011.

[19] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Computational Learning Theory*, 1995.

[20] N. C. Oza and S. Russell, "Online bagging and boosting," in *Artificial Intelligence and Statistics*, 2001, pp. 105–112.

[21] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, vol. 1, pp. 511–518, 2001.

[22] S. Avidan, "Support vector tracking," *IEEE Trans. on Pattern Analysis and Machine Intell. (PAMI)*, vol. 26, no. 8, 2004.

[23] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, 1979.

[24] I. J. Cox and S. L. Hingorani, "An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 18, no. 2, pp. 138–150, 1996.

[25] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, 2008.