On Multi-Modal People Tracking from Mobile Platforms in Very Crowded and Dynamic Environments

Timm Linder[1]   Stefan Breuers[2]   Bastian Leibe[2]   Kai O. Arras[1,3]

***Abstract*— Tracking people is a key technology for robots and intelligent systems in human environments. Many person detectors, filtering methods and data association algorithms for people tracking have been proposed in the past 15+ years in both the robotics and computer vision communities, achieving decent tracking performances from static and mobile platforms in real-world scenarios. However, little effort has been made to compare these methods, analyze their performance using different sensory modalities and study their impact on different performance metrics. In this paper, we propose a fully integrated real-time multi-modal laser/RGB-D people tracking framework for moving platforms in environments like a busy airport terminal. We conduct experiments on two challenging new datasets collected from a first-person perspective, one of them containing very dense crowds of people with up to 30 individuals within close range at the same time. We consider four different, recently proposed tracking methods and study their impact on seven different performance metrics, in both single and multi-modal settings. We extensively discuss our findings, which indicate that more complex data association methods may not always be the better choice, and derive possible future research directions.**

I. Introduction

People tracking from a first-person perspective using a mobile sensor platform has been studied in the robotics and computer vision communities for over a decade, and various detection methods for different sensor modalities, as well as tracking algorithms have been proposed. Often, complex, generic data association methods are combined with extensions that are specific to the application domain to better deal with frequent occlusions in such environments and model people's behavior [1], [2]. However, very recent work [3] reminds us that although complex data association methods, such as JPDAF or MHT, have been shown to deliver better performance in application areas with high-clutter environments like radar tracking [4], no systematic comparison between simpler and more complex data association methods has been performed for people tracking, where false positive detections occur systematically rather than randomly. Also, most systems focus only on a single sensor modality, and are tested in simple environments with only few tracked persons and limited dynamics.

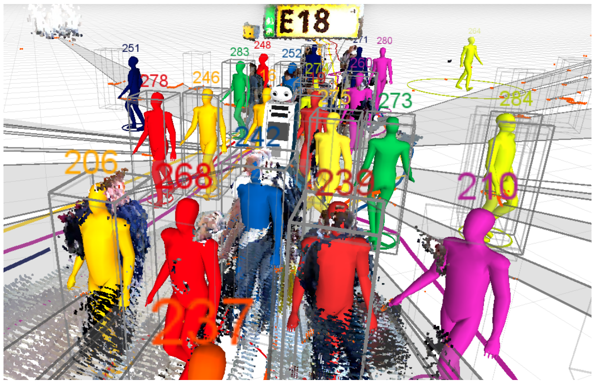


Fig. 1. Typical example of a crowded, dynamic situation in an airport terminal with frequent occlusions where we want to robustly and efficiently track persons from a first-person perspective with our mobile service robot platform, which can (barely) be seen in the center of the picture.

In this paper, we want to go one step further and examine how well some recent, publicly available tracking methods perform in challenging, highly crowded and dynamic scenarios such as a busy airport terminal (Fig. 1). Following recent trends in the computer vision community towards a standardized benchmark for multi-object tracking methods [5], [6], and to enable a fair comparison of different tracking methods, we integrate them into a common framework and provide them with the same set of detections as input.

Our contributions are:

- An extensive ROS-based framework that provides the tooling for the systematic evaluation of multi-modal people tracking algorithms under identical conditions
- A comparison, with focus on both tracking quality and runtime performance, of four state-of-the-art real-time tracking systems [7]–[10] that have been integrated into the framework – including a proven MDL-based tracking approach from the computer vision community
- Experiments on two challenging new datasets with RGB-D and 2D laser from a first-person perspective
- A thorough discussion of strengths and weaknesses of current methods in these scenarios, and possible future directions of research.

Large parts of our framework, including our new multi-modal annotation tool and the parameters used to obtain our results, are publicly available as open source, to allow researchers to quickly reproduce our results on their own datasets and to evaluate their own algorithms using our framework.

II. Related work

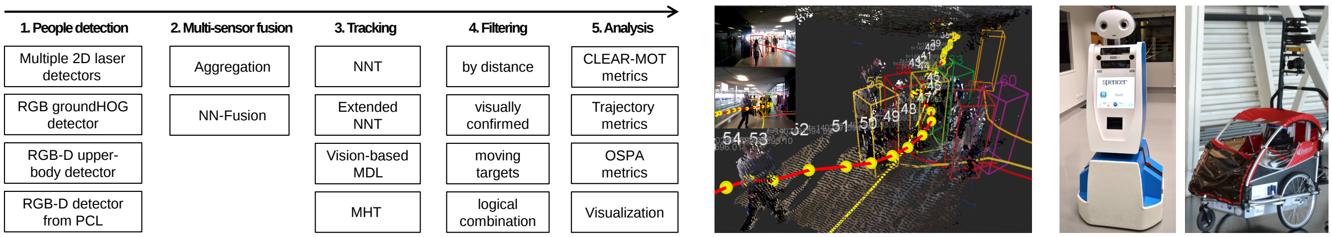
People detection and tracking are of high interest to both robotics and computer vision. While both communities have,

[1] Social Robotics Laboratory, University of Freiburg, Germany. http://srl.informatik.uni-freiburg.de, {linder,arras}@cs.uni-freiburg.de

[2] Visual Computing Institute, RWTH Aachen University, Germany. http://www.vision.rwth-aachen.de/, {breuers,leibe}@vision.rwth-aachen.de

[3] Bosch Corporate Research, Renningen, Germany

Both first authors contributed equally to this work, which has been partly supported by the European Commission under contract number FP7-ICT-600877 (SPENCER).



(a) Our modular, multi-modal tracking framework (b) Multi-modal annotation tool (c) Robot (d) Sensor platf.

Fig. 2. (a) All components of our framework are implemented as separate, reusable ROS modules, most of them open source. So far, we have integrated four existing tracking methods [7]–[10] into our framework for a fair comparison under identical conditions. (b) Our new multi-modal track annotation tool, based upon *RViz*. The 3D visualization encompassing RGB-D and laser point clouds as well as annotated trajectories (red line with yellow waypoint markers), along with 2D camera views with projected annotations (small images), significantly speeds up the annotation process. (c) Our service robot platform, equipped with front- and rear-facing 2D laser and RGB-D sensors. (d) Our mobile sensor platform, in a similar sensor configuration.

rather individually, made significant progress in the past, there has recently been a trend towards combining multiple methods and modalities as the available computational power for real-time detection and tracking is becoming larger, also on mobile robot platforms.

In robotics, often laser sensors are used to cover a large field of view, especially for mapping and navigation. The methods to detect people in this setup are based on simple ad-hoc classifiers looking for local minima in the scan [11], [12], or more elaborate person detectors [13]. In vision, camera and RGB-D sensor information is used as an input stream to the detection pipeline. Often HOG features are used to detect full bodies [14], [15], while upper-body detectors like [10] are better suited to detect nearby persons.

In both areas, missed detections and false alarms need to be compensated by a tracking algorithm, which often uses some form of data association method. [8] and [16] use the most simple NN method or NN-JPDA. Another NN tracker, [9], is using a more sophisticated track initiation and deletion logic and an interacting multiple model filter (IMM). The more complex multi-hypothesis tracking methods [17], [18] are known to outperform simple NN methods in radar tracking [4], and have also been used for people tracking purposes [7] along with extensions such as a social force model or person-level feedback from group-tracking [1], [19]. The vision based MDL-tracker [10] is also loosely based on MHT, but formulates it as a quadratic pseudo-boolean optimization problem, solved via mininum description length (MDL). As it was specifically designed for visual data, it allows for the incorporation of an appearance model. Most trackers use an Extended Kalman Filter (EKF) to incorporate a constant velocity motion model of pedestrians.

With higher computational power, it has become possible to use multi-modal sensor platforms, equipped with both laser and RGB-D sensors, especially on service robots. This combination makes it possible for the robots to deal with the challenges in their highly crowded and dynamic field of operation. While a few multi-modal systems have been presented in the past [8], [20], [21], to the best of our knowledge, a consistent comparison of different people tracking approaches in a multi-modal setup in challenging environments is still missing in robotics. Even in the vision community, a standardized baseline evaluation of existing tracking methods has just begun [6]. Being aware of the challenges of groundtruth evaluation, as discussed in [5], we aim at providing a reusable, multi-modal framework that enables a consistent, comparative evaluation.

III. Our framework

Fig. 2a gives an overview of the main components of our modular people tracking framework. All of these components are fully integrated into ROS and publicly available and documented on GitHub[1]. In the following, we will briefly describe the most important components, starting from the left at the detection layer.

A. Detection

2D laser. For people detection in 2D laser range data, we use a random forest classifier trained on the laser features described in [13]. While our ROS-based implementation, using classifiers from the OpenCV library, also allows to use other classifiers such as Adaboost or SVM, the random forest (with 15 trees and maximum depth of 10) performed best on our manually annotated training/test data set recorded in a pedestrian zone [19] using a SICK LMS 500 laser scanner at a height of 70 cm and 0.25 degrees angular resolution. After a separate ROS node has segmented the laser scans using a variant of jump-distance clustering, the detector computes a set of geometric 2D features on each segment which are then fed to the classifier.

Monocular vision and RGB-D. For person detection in RGB-D, the existing depth template-based upper-body detector described in Jafari et al. [10], which runs in real-time at 20-30 Hz on the CPU, as well as their CUDA-based, monocular groundHOG detector [14] have been integrated. We also extended the RGB-D person detector from [22], which applies a HOG classifier on candidate regions extracted from a depth-based height map, with GPU acceleration.

Fusing detections. For multi-sensor people tracking, our framework allows to flexibly combine detections from multiple modalities using a detection-to-detection fusion scheme,

[1] `https://github.com/spencer-project/spencer_people_tracking`

easily set up via XML, that works even when the particular tracking algorithm was not specifically designed to cope with detection input from multiple sources, as most of the trackers in our evaluation. Using greedy NN association, we first fuse detections from sensors with overlapping fields of view (*e. g.* front laser, front RGB-D) and then aggregate the resulting sets of detections that do not overlap (*e. g.* front and rear detections). As association cost, we use either the Euclidean or Mahalanobis distance between individual detections, or a cost computed in polar coordinates that penalizes discrepancies in distance less heavily (mainly useful for 2D image-based detectors that do not output precise depth estimates).

All detectors integrated into our framework output detections which adhere to the same ROS message format. A *DetectedPerson* comprises a position vector $\mathbf{z}'$ and its uncertainty R' in a sensor-specific 3D coordinate frame, a scalar detection confidence, and some meta-data. This clearly defined interface allows to easily integrate additional detectors into the system, and provides the interface to the tracking module.

B. Tracking

Up to now, we have integrated four different tracking systems into our framework for comparison purposes. We will shortly outline these approaches in the following:

Nearest-neighbor tracker [8]. This is a very fast tracker based upon a nearest-neighbor data association that has recently been integrated into ROS by the authors of [8]. Motion prediction is performed via an Extended Kalman Filter (EKF) using a constant velocity (CV) motion model, and tracks are initiated if a minimum number of detections occur within a small radius. Track deletion takes place if the track covariance exceeds a certain limit. A more advanced NN-JPDAF association method, not used in our experiments, is also provided.

Extended nearest-neighbor tracker [9]. Also based upon greedy NN data association, this recent work of our own was developed with robustness and computational efficiency in mind especially in highly crowded scenarios. Compared to [8], it includes a velocity-based track initiation logic that only initiates new tracks if a given amount of detections appear with a consistent velocity profile that is compatible with typical human walking speeds. Track deletion occurs when the number of tracking cycles without matching detection exceeds a certain threshold, and a distinction is made between 'young' and 'mature' tracks that have already existed for a while; young tracks are deleted after a smaller number of cycles, since they often represent false alarms. For motion prediction, an IMM approach is used that combines multiple CV and coordinated turn models.

Multi-hypothesis tracker [7]. As an additional baseline method for comparison, we use a variant of the multi-hypothesis tracker (MHT) after Reid *et al.* [17] and Cox & Hingorani [18] with explicit occlusions labels [7]. This probabilistic tracker does not incorporate any track initiation logic. Instead, new track creation is modelled via a Poisson process. Various extensions (such as incorporation of group-level feedback or a social force model) have been proposed in the past, but are not used in our experiments. For the purpose of tracking people from a moving platform in dynamic environments, the MHT is configured with a low scanback depth to enable real-time decision making at low latencies.

Vision-based MDL tracker [10]. This method is based on the work of Leibe *et al.* [23] and uses the tracking framework of [24], [25] to build an overcomplete set of track hypotheses, similar to MHT. Via bi-directional EKF new trajectories are generated for the current frame by following the motion model backwards in time, while existing trajectories are extended from the last to the current frame. Each track then receives an individual score incorporating the motion model as well as confidence and color-based appearance of inlier detections, adapted on the fly. The interaction cost between tracks takes into account physical overlap and shared detections. Selecting the best subset of hypotheses from the score matrix is then formulated as a quadratic binary problem and solved in an MDL fashion by the multi-branch method of [26].

In all of the examined tracking systems, person detections arrive in their sensor-specific coordinate frame and are instantaneously transformed into a locally fixed frame (based upon robot odometry) that does not move with the robot. This ensures that the motion prediction of tracked persons is independent from the robot's ego-motion. In the resulting set of measurements $Z = \{\mathbf{z}_1, ..., \mathbf{z}_n\} \subset \mathbb{R}^2$, we drop the z coordinate as we only track in 2D world coordinates.

C. Groundtruth annotation and evaluation

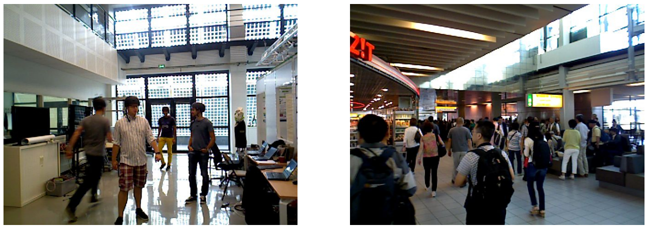
To our knowledge, no multi-modal track annotation tool for 2D/3D laser, RGB-D and stereo data is currently publicly available. Our new ROS-based multi-modal annotation tool, partially shown in Fig. 2b, leverages the powerful visualization capabilities of the ROS visualization tool *RViz* and *rqt* to enable annotating people directly in 3D world space in the RGB-D and laser point clouds. For reference purposes, annotations and 2D laser scans are also projected into the camera images. By placing trajectory waypoints at regular intervals (e. g. every 0.5 sec or 2.0 sec, depending on the dynamics of the scene) and interpolating in between, the annotation process is significantly sped up.

For tracking performance evaluation purposes, we have integrated and extended a publicly available Python implementation of the CLEAR-MOT metrics, and implemented further trajectory-based metrics (both described in Sec. IV-B), as well as the OSPA metrics [27] (not used in this work).

IV. Experiments

A. Datasets

For our experiments, we have recorded two entirely new, multi-modal datasets (cf. Fig. 3). The *Motion Capture Sequence* has been recorded in a narrow lab environment in front of our robot platform, shown in Fig. 2c, which



(a) Motion Capture Sequence (b) Airport Sequence 03

Fig. 3. Example RGB frames from our new datasets.

remains stationary throughout this sequence. The recorded sensor data includes a 190-degree frontal 2D laser scan from a SICK LMS 500 sensor, and the data from a front-facing Asus Xtion Pro Live RGB-D sensor. In this four-minute sequence, available via our website, four persons that wear motion capture markers on their heads for groundtruth acquisition are moving around in highly dynamic and erratic patterns, very frequently occluding each other and stopping or accelerating abruptly. This dataset is mainly intended to simulate human-robot interaction, in which case people often 'play' with the robot, or try to challenge its tracking abilities.

The second sequence, *Airport Sequence 03*, is part of a much larger dataset that was recorded at Amsterdam-Schiphol airport, used by around 150,000 passengers each day, using a moving sensor platform (Fig. 2d) that closely replicates the sensory setup on our robot. The dataset includes 2D laser range data recorded from two back-to-back SICK LMS 500 scanners at 70 cm height, covering a *full 360-degree horizontal field of view* around the robot. It also includes RGB-D data from two Asus sensors mounted in horizontal orientation and facing into forward and rearward direction.

In the first half of this sequence, the sensor platform remains stationary and observes a dense flow of passengers disembarking from an airplane. In the second half, the platform joins the flow of people towards a large, open area inside the terminal. During the entire 4-minute sequence, the platform is almost constantly surrounded by 20–30 persons that follow various motion patterns at different walking speeds and undergo many severe occlusions. In total, 172 ground truth tracks have been manually annotated using our new multi-modal annotation tool. For our experiments, we ignore all groundtruth tracks at a distance of greater than 12.0m as correctly annotating tracks becomes highly challenging above this distance due to extreme occlusions and increasing inaccuracy in sensor calibration.

B. Evaluation metrics

A commonly used measure for evaluating multi-object tracking performance is the CLEAR-MOT metrics [28]. Besides counting false positives (FP), false negatives (FN) and ID switches (IDS), they define an aggregate error measure called MOT Accuracy (MOTA) as

$$\mathrm{MOTA} = 1 - \frac{\sum_k (\mathrm{FP}_k + \mathrm{FN}_k + \mathrm{IDS}_k)}{\sum_k \mathrm{GT}_k},$$

where k is the tracking cycle index. The optimal MOTA score is 1.0, and MOTA can reach negative values if the tracker makes more errors than there are ground truth objects GT over the entire duration of the dataset.

As discussed extensively in [5], MOTA scores can vary between implementations and are highly dependent on meta-parameters such as the matching distance threshold θ_d and the way in which track hypotheses are assigned to groundtruth objects. In our version, we compute groundtruth correspondences using a variant of the Hungarian method, based upon Euclidean distances between object centroids in world coordinates with $\theta_d = 1\mathrm{m}$. We ignore all correspondences where the groundtruth track is physically occluded, which is determined by searching for associated laser points within a radius of $0.3m$ of the annotated position, shifted towards the sensor origin by $0.2m$ to take into account that the laser sensor only perceives the surface of the person.

One caveat is that the number of ID switches (IDS) has very low influence on overall MOTA, as FP and FN counts are often significantly higher in comparison. Therefore, the absolute number of ID switches is often used as a second, separate measure when evaluating people tracking performance. However, a tracking system with lower track recall (i. e. which tracks less persons by, for instance, initiating tracks very reluctantly) almost certainly generates less ID switches. Very recent research in the computer vision community [6], which we adopt here, instead motivates to compute the *relative number of ID switches*, rIDS, defined as a product of the absolute number of ID switches and the inverse of the recall over all frames, $\mathrm{IDS} \cdot \mathrm{GT}/\mathrm{TP}$.

Finally, we also compute the trajectory-based measures of mostly tracked (MT) and mostly lost (ML) persons [29], denoting the number of groundtruth tracks that have been tracked for more than 80% or less than 20% of their length.

C. Experimental setup

All of our experiments were conducted on a high-end gaming laptop equipped with a quad-core Intel Core i7-4700 MQ processor and 8 GB of RAM under Ubuntu 14.04 with ROS Indigo. Each single experiment has been run at least 3 times and metrics have been averaged to ensure stable results that are not negatively affected by the not fully deterministic message passing, synchronization and transform lookups in ROS. For the computationally more complex experiments on the airport dataset sequence, we have pre-recorded all detections to ensure that all tracking algorithms are always fed with the same input for a fair comparison.

D. Parameter selection

As highlighted in [5], for evaluations of multi-person tracking it is important that parameters of the tracking algorithm are tuned on a separate validation dataset to verify its generalization capabilities and avoid overfitting. With this in mind, we carefully tuned all of our algorithms on separate, similar, but not identical datasets. Specifically, for tuning the NNT [8], the Extended NNT [9] and the MHT [7], we used the laser-based *Freiburg Main Station* dataset (cf. e. g. [1]), as well as a synthetically generated dataset via a combination of the pedestrian simulator *PedSim* and *Gazebo* (see [9] for details). The vision-based MDL tracker has been tuned on the *ETH* dataset [30] recorded in a pedestrian zone.

V. Results

In Tables I–IV, we present quantitative results of the different tracking methods for each modality on our datasets. Qualitative results are available on our YouTube channel[2]. As the MDL tracker [10] currently only supports image-based detections from the upper-body and groundHOG detectors, we only use it in experiments with the front RGB-D sensor.

A. *Comparison of different tracking approaches*

Comparing the results of the different tracking approaches under identical conditions (sequence, modality), we note that the simple NN approaches often generate the best MOTA score. This might be due to a lower number of parameters, which could result in better generalization capabilities regarding new scenarios. The Extended NNT is superior to the NNT in terms of MOTA and FP%, most likely due to its track initiation and deletion logic. Especially on the *Motion Capture Sequence* with four groundtruth tracks, one or two ghost tracks are enough to cause bad FP scores for methods without a sophisticated initiation logic, such as NNT and MHT. Interestingly, both of these perform very similarly in most of the tested scenarios concering MOTA and FP%. On the other side, the miss ratio is often the highest for the Extended NNT, and caused by delayed track initiation.

Both multi-hypothesis methods seem to suffer from frequent switching between hypotheses, a problem well known in multi-hypothesis tracking. This results in a high number of relative ID switches. However, in front RGB-D only (Tab. I), the MDL-Tracker gives best FP% and thus a MOTA score comparable to the one of Extended NNT. Note that MHT might obtain better results if parameters such as new track rates were re-tuned on the datasets used for testing, or if it were allowed to delay decision making by backtracking in a fixed-lag smoothing sense. Nevertheless, this can be problematic for real-time motion planning applications due to the introduced delay, and lowers MOTA when comparing always against the most recent groundtruth.

The simple NNT tracker dominates in the number of consistently tracked targets, i.e., higher MT and lower ML, due to a more straightforward initiation of tracks.

B. *Laser-only vs. multi-modal detections*

Next, we want to investigate the benefits of the multi-modal sensor platform and the use of both 2D laser and RGB-D sensors. On the airport sequence, incorporating vision-based detections from groundHOG and upper-body increases the number of mostly tracked targets. This leads to a higher track recall and lower miss ratio for all approaches, at the cost of an increased FP%, ultimately resulting in a lower MOTA score (Tab. III). A more sophisticated fusion scheme of the different detector outputs might yield some improvement, however, a visual inspection of our naïve fusion scheme reveals no immediately apparent problems. Instead, further experiments reveal that the HOG detector causes many false alarms (Fig. 4) and provides imprecise depth estimates for distant persons, obtained by projecting image footpoints onto the estimated ground plane. Tab. IV shows the multi-modal result without HOG, but still using upper-body detections from the RGB-D sensor. The FP% decreases, but unfortunately also MT goes down and miss rate increases. Anyhow, general tracking quality improves, which is reflected in the highest MOTA score for each tracking approach so far using this configuration.

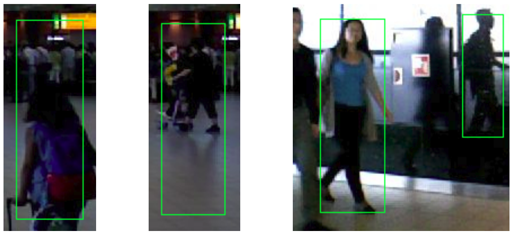

(a) (b) (c)

Fig. 4. Typical failure cases from the HOG detector caused by (a) clutter, (b) reflections on the floor or (c) on the walls. Reflections on walls and glass surfaces can sometimes also cause false laser detections.

On the *Motion Capture Sequence*, all methods struggle with an extremely high FP%, except for Extended NNT, whose extensive track initiation logic can again compensate for false alarms. The resulting discrepancy of the MOTA scores is huge (75% vs. 8-15%). It seems here that the laser detector – instead of HOG – is responsible for most of the false positives, often in chairs and other furniture: when using only front RGB-D (Tab. I), FP% is around 50%-points lower.

C. *Filtering detections by a static map*

As a static map of the environment is often available for navigation purposes anyway, we want to examine its use for false positive suppression. We rasterize a circle of 15 cm radius at the detection's position onto the given occupancy grid map. If less than 90 percent of all grid cells are free, we reject the detection. As we filter on the detection level, the process can be applied to any detector.

As no map had been recorded in the airport environment, we restrict this experiment to the *Motion Capture Sequence* (Tab. V), where it leads to an increase in MOTA of 15–65 percentage points for the different tracking approaches.

D. *Runtime performance*

In the last column of each table, we show the median of the extrapolated processing rates of the tracking algorithms based upon actually measured cycle times (without taking the detection stage into account). All examined tracking systems are implemented in C++. Note that the rate of MHT is fixed to 30 Hz, generating as many hypotheses per cycle as possible in this time frame (at lower rates, the performance gets worse due to less frequent updates of the EKF).

[2] https://youtube.com/spencereuproject

	Airport Sequence 03							**Motion Capture Sequence**						
Method	MOTA	rIDS	FP%	Miss%	MT	ML	Hz	MOTA	rIDS	FP%	Miss%	MT	ML	Hz
NNT [8]	27.7%	227	39.4%	**32.5%**	**92**	**47**	**13701**	60.7%	**131**	23.6%	**14.3%**	4	0	**20726**
Extended NNT [9]	**44.4%**	**210**	13.1%	42.1%	63	60	4287	**69.8%**	151	7.8%	20.9%	4	0	5637
MHT [7]	26.9%	338	39.4%	33.0%	87	51	28	57.9%	173	24.7%	15.6%	4	0	28
MDL-Tracker [10]	43.7%	428	**12.5%**	43.1%	36	59	53	60.7%	373	**4.8%**	31.3%	1	0	139

TABLE I

Only front RGB-D detections (small FOV)

	Airport Sequence 03							**Motion Capture Sequence**						
Method	MOTA	rIDS	FP%	Miss%	MT	ML	Hz	MOTA	rIDS	FP%	Miss%	MT	ML	Hz
NNT [8]	59.7%	**236**	20.8%	**19.3%**	**112**	27	**6184**	25.6%	**54**	70.4%	**3.3%**	4	0	**17968**
Extended NNT [9]	**62.8%**	331	**3.4%**	33.5%	68	35	2307	**68.8%**	60	**25.3%**	5.2%	4	0	4988
MHT [7]	58.9%	700	16.6%	23.9%	85	**26**	29	28.0%	85	67.3%	3.8%	4	0	28

TABLE II

Only laser detections (large FOV)

	Airport Sequence 03							**Motion Capture Sequence**						
Method	MOTA	rIDS	FP%	Miss%	MT	ML	Hz	MOTA	rIDS	FP%	Miss%	MT	ML	Hz
NNT [8]	45.7%	325	36.4%	**17.7%**	**123**	**19**	**4590**	14.8%	**55**	81.7%	**2.7%**	4	0	**15703**
Extended NNT [9]	**62.1%**	**313**	**8.2%**	29.4%	96	26	2005	**74.9%**	58	**20.1%**	4.3%	4	0	4690
MHT [7]	46.3%	692	34.9%	18.2%	117	22	31	8.6%	74	87.6%	2.9%	4	0	29

TABLE III

Multi-modal detections (large FOV)

	Airport Sequence 03							**Motion Capture Sequence**						
Method	MOTA	rIDS	FP%	Miss%	MT	ML	Hz	MOTA	rIDS	FP%	Miss%	MT	ML	Hz
NNT [8]	62.1%	**226**	18.7%	**19.0%**	**114**	27	**6100**	18.1%	**52**	77.6%	3.7%	4	0	**15857**
Extended NNT [9]	**64.2%**	262	**3.3%**	32.4%	77	33	2222	**77.4%**	62	**16.5%**	5.4%	4	0	4744
MHT [7]	60.2%	676	17.2%	22.0%	97	**24**	29	17.8%	76	77.7%	**3.6%**	4	0	28

TABLE IV

Multi-modal detections without HOG (large FOV)

The simple NNT is about 3 times computationally more efficient than the Extended NNT. Both outperform the two more complex methods MDL and MHT by two orders of magnitudes, and require less than 10 percent CPU usage even in very crowded environments.

VI. Discussion

In the following, we discuss the most important conclusions that we can draw from our experiments on the crowded airport dataset and the dynamic motion capture sequence.

A. *Influence of detector performance*

A major observation that we made during our experiments is that detector performance is the single, most important factor influencing tracking performance which goes far beyond the impact of the chosen tracking algorithm. In a nutshell, none of the examined tracking methods deal really well with high false positive rates. During initial experiments, we used a 2D laser detector trained on a different sensor model, and using a less restrictive selection of positive and negative training samples. This detector caused extremely bad MOTA scores between -3.3 and -1.3 due to enormous false positive rates (> 200%). None of the examined methods could cope with this high number of false positives, which occur systematically and repeatedly at the same locations. Although the track initiation logic of the extended NNT was able to suppress a significant amount of the false positives, MOTA still did not exceed -1.3. Even though multi-hypothesis trackers have been shown to work well in (random) high clutter in radar tracking [4], the worst MOTA score was obtained using the MHT, which does not possess any dedicated track initiation logic. After incorporating the static occupancy grid map to filter out false detections beforehand, MOTA scores of all examined approaches became positive, but were still significantly below the levels presented in Sec. V.

B. *Integrating 2D image-based detections*

While a vision-based tracker can significantly benefit from 2D image-based detections (e. g. from HOG) that extend its

	Motion Capture Sequence				
Method	MOTA	rIDS	FP%	Miss%	Hz
NNT [8]	78.0%	**50**	18.5%	2.9%	**19050**
Extended NNT [9]	**89.4%**	59	**5.3%**	4.6%	4926
MHT [7]	73.8%	71	21.9%	3.4%	28

TABLE V

MULTI-MODAL DETECTIONS (LARGE FOV) + STATIC MAP

maximum tracking distance beyond the useful working range of RGB-D sensors (around 6–7m), their depth estimates are often very imprecise. In a multi-modal setup where precise laser measurements are available ($\sigma \approx$ 3cm), using HOG detections as a direct input to the tracking algorithm may therefore be detrimental. Instead, laser-based detections may be a better choice to cover far detection ranges, while image-based detections could be used to validate laser detections visually, if an association can be established.

C. The choice of tracking parameters

Our experience shows that the correct choice of parameters significantly outweighs the choice of data association. Tracking approaches with only few parameters may generally be the preferrable choice, as they may generalize better towards new scenarios. Especially complex, probabilistic multi-hypothesis approaches often require re-learning or manual tuning by an expert of parameters such as new-track or deletion Poisson rates that depend on the given scenario and can vary with location and time (e. g. when a new plane arrives at an airport and the passengers start disembarking). Also, automatic parameter learning approaches as outlined in [9], [31] may help to simplify the process. To make our results easier to reproduce and allow researchers from other fields (e. g. HRI) to benefit from our findings, we will share all parameter configurations used in our experiments online.

D. Trade-off between FPs, miss rate and ID switches

Another lesson we learn from our experiments is that the choice of parameters greatly depends on the desired application scenario. There appears to be no universal set of parameters that fully accommodates all requirements, as a trade-off has to be made between attaining a low false positive count, a low miss rate, and a low number of ID switches. The first two can be important when using the people tracker output for socially aware navigation, since high false positive rates (i. e. ghost tracks) could freeze the robot, while missed tracks can cause the robot to behave impolitely or even endanger people. On the other hand, in person guidance scenarios, it is of utmost importance to maintain the ID of a tracked person as long as possible, while false positives might not be such a large issue.

As shown in our experiments, low false positive rates can be achieved by a dedicated track initiation logic, pre-filtering on a static map, and early track deletion. The first two options can cause the tracker to miss certain (e. g. static) tracks, while the last option may result in ID switches if the track suddenly reappears after an occlusion.

E. Importance of standardized tracking metrics

Even minor differences in tracking metrics implementation or its parameters can have significant influence on results. We agree with findings from the vision community [5] which underline the importance of using a standardized evaluation script and the same detection input for all tracking systems. Our proposed tracking framework is, to our knowledge, the first that allows for *multi-modal* data annotation in RGB-D, 2D/3D laser and potentially stereo data, and enables a systematic evaluation and comparison of different tracking methods and detectors in a joint framework.

F. Which tracking approach to choose?

Finally, we attempt to answer the question which of the examined tracking methods to choose for real-time people tracking from a mobile platform in very crowded and dynamic environments. Looking at the multi-modal results on the motion capture sequence (Table III, right), we see that the same, underlying NN data association method of [8], [9] delivers an astonishing difference in MOTA performance of 60%, depending on the presence or lack of a dedicated track initiation logic. On the other hand, on the airport sequence, we observe only a a 0.5% difference in MOTA between simple NN and complex MHT data association. Therefore, as already hinted at in [9], it appears that incorporating promising tracking extensions (e. g. [1], [2]) into a simple data association scheme might be the way to go. The computation time which is saved by refraining from using a more complex multi-hypothesis data association method could instead be spent on higher-level perception, or to improve detector performance, which has a high impact as previously discussed. Both of the discussed NN-based approaches are relatively easy to configure, show good performance on our test datasets, and run at low CPU usage (<10% on a single core) – which is crucial on a mobile service robot platform that also needs to localize itself, plan and navigate.

Here, the hypothesis-oriented MHT after Reid [17] and Cox & Hingorani [18] may also be at disadvantage in very crowded environments. Since the entire state of the scene is represented within each single hypothesis, a very large number of hypotheses may be needed to adequately represent all likely combinations of possible track states. In [7], up to 1000 hypotheses are generated for just 4 person tracks, each one involving the same data association that the NN-based methods only need to perform once. Generating as many hypotheses as possible within a given time window, as in our experiments, ensures a certain minimum cycle rate to be met, but may result in only few hypotheses being generated.

Finally, scenarios where some delay in decision making can be tolerated, such as offline video analysis or static observation of people behavior, allow for a different mode of evaluation where the delayed selection of the best hypothesis can be taken into account, by deferring matching with the groundtruth by a certain number of tracking cycles. We believe that in these cases, the multi-hypothesis approaches [7], [10] can show their full potential and attain higher scores.

G. Future directions of research

Using solely detectors with relatively low false-positive rate, the difference in tracking metrics between various tracking approaches and implementations becomes surprisingly small. Visually analyzing the remaining ID switches that still occur on the airport and motion capture sequences, we believe that in these cases, the motion model provides insufficient information and full person reidentification is required. Implementing a robust reidentification module can be very challenging in scenarios such as the airport, which is used by over 150,000 passengers per day. An open question is still how to deal with tracks that first re-appear in 2D laser and need to be assigned a preliminary ID, before potentially getting visually re-identified as a previously seen person; in person guidance scenarios, this issue potentially needs to be dealt with on the task planning level.

VII. Conclusion

In this paper, we have presented a multi-modal people tracking framework into which we have integrated four existing tracking approaches of varying complexity, in order to study them on two challenging new, multi-modal datasets – one of them recorded from a static platform in a highly dynamic HRI scenario, and another one from a moving platform inside a crowded airport terminal. We have carefully analyzed the performance of these existing methods with regard to multiple tracking metrics under different multi-modal configurations, identified and extensively discussed their strengths and weaknesses, shared some learned lessons and drawn conclusions that may guide possible future directions of research. Finally, we want to encourage other researchers to integrate their own detectors and tracking algorithms into our framework, and share their results.

References

[1] M. Luber and K. O. Arras, "Multi-hypothesis social grouping and tracking for mobile robots," in *Proceedings of Robotics: Science and Systems*, 2013.

[2] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras, "People tracking with human motion predictions from social forces," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

[3] A. Leigh, J. Pineau, N. Olmedo, and H. Zhang, "Person tracking and following with 2d laser scanners," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[4] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*, 1999.

[5] A. Milan, K. Schindler, and S. Roth, "Challenges of ground truth evaluation of multi-target tracking," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, 2013.

[6] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1504.01942

[7] K. O. Arras, S. Grzonka, M. Luber, and W. Burgard, "Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2008.

[8] C. Dondrup, N. Bellotto, F. Jovan, and M. Hanheide, "Real-time multisensor people tracking for human-robot spatial interaction," in *Workshop on Machine Learning for Social Robotics at International Conference on Robotics and Automation (ICRA)*, 2015.

[9] T. Linder, F. Girrbach, and K. O. Arras, "Towards a robust people tracking framework for service robots in crowded, dynamic environments," in *Assistance and Service Robotics Workshop (ASROB-15) at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[10] O. H. Jafari, D. Mitzel, and B. Leibe, "Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[11] A. Fod, A. Howard, and M. Mataríc, "Laser-based people tracking," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2002.

[12] 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, 2003.

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

[14] P. Sudowe and B. Leibe, "Efficient use of geometric constraints for sliding-window object detection in video," in *Computer Vision Systems*, 2011.

[15] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2011.

[16] N. Bellotto and H. Hu, "Computationally efficient solutions for tracking people with a mobile robot: an experimental evaluation of bayesian filters," *Autonomous Robots*, vol. 28, no. 4, 2010.

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

[18] I. Cox and S. Hingorani, "An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Trans. on Pattern Analysis & Machine Intell.*, vol. 18, no. 2, 1996.

[19] T. Linder and K. O. Arras, "Multi-model hypothesis tracking of groups of people in RGB-D data," in *IEEE Int. Conf. on Information Fusion (FUSION'14)*, 2014.

[20] E. Schaffernicht, C. Martin, A. Scheidig, and H.-M. Gross, "A probabilistic multimodal sensor aggregation scheme applied for a mobile robot," in *KI 2005: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, 2005, vol. 3698.

[21] L. Spinello, R. Triebel, and R. Siegwart, "Multiclass multimodal detection and tracking in urban environments," *The International Journal of Robotics Research*, vol. 29, no. 12, 2010.

[22] M. Munaro, F. Basso, and E. Menegatti, "Tracking people within groups with RGB-D data," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012.

[23] 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 & Machine Intell.*, vol. 30, no. 10, 2008.

[24] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "Robust multiperson tracking from a mobile platform," *IEEE Trans. on Pattern Analysis & Machine Intell.*, vol. 31, no. 10, 2009.

[25] K. Schindler, A. Ess, B. Leibe, and L. Van Gool, "Automatic detection and tracking of pedestrians from a moving stereo rig," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, 2010.

[26] K. Schindler, U. James, and H. Wang, "Perspective n-view multibody structure-and-motion through model selection," in *Proc. of the European Conf. on Comp. Vision (ECCV)*, 2006.

[27] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *Signal Processing, IEEE Transactions on*, vol. 59, no. 7, 2011.

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

[29] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybrid-boosted multi-target tracker for crowded scene," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[30] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[31] M. Luber, G. D. Tipaldi, and K. O. Arras, "Better models for people tracking," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2011.