Multi-Model Hypothesis Tracking of Groups of People in RGB-D Data

Timm Linder Kai O. Arras

Social Robotics Lab, University of Freiburg
79110 Freiburg, Germany, {linder, arras}@cs.uni-freiburg.de

***Abstract*—Detecting and tracking people and groups of people is a key skill for intelligent vehicles, interactive systems and robots that are deployed in humans environments. In this paper, we address the problem of detecting groups of people from learned social relations between individuals with the goal to reliably track group formation processes. Opposed to related work, we track and reason about multiple social grouping hypotheses in a recursive way, assume a mobile sensor that perceives the scene from a first-person perspective, and achieve good tracking performance in real-time using RGB-D data. In experiments in large-scale outdoor data sets, we demonstrate how the approach is able to track groups of people with varying sizes over long distances with few track identifier switches.**

***Index Terms*—Service robots, robot sensing systems, computer vision, social factors**

I. Introduction

Tracking groups of people is an important skill for surveillance systems, intelligent vehicles and robots that operate in populated environments. Empirical research has found that up to 70% of pedestrians walk in groups [1]. Knowledge about groups, their position, size, motion state, and social activities can enable systems to gain a deeper understanding of human environments and to provide better services to users. Examples include multi-party human-computer interaction or socially compliant robot navigation among groups of people. The goal in this work is tracking groups of people from a mobile sensor in a first-person (within-scene) perspective. Unlike stationary overhead cameras in surveillance applications, this is a challenging scenario because with sensors at human eye-level, people in groups are occluded more frequently and are harder to detect and track reliably as individual targets. In addition to the estimates on social relations between people, group tracking also allows to improve person-level tracking by feeding back the grouping information, for example, to better deal with lengthy occlusions of individual targets.

The problem of tracking groups of people has been addressed using image data as well as 2D range data from on-board laser scanners. To our knowledge, this work is the first to use RGB-D data to this end. We extend an approach by Luber *et al.* [2] based on a multi-model multi-hypothesis group tracker with a mechanism to maintain consistent group IDs across multiple group splits and merges. The approach tracks group formation processes explicitly and allows for the representation of multiple social grouping hypotheses at the same time. This is achieved by extending a multiple hypothesis tracker (MHT) by an alternating model hypothesis step. We conduct experiments on unscripted, real-world data captured with a mobile platform equipped with two RGB-D sensors in a crowded pedestrian zone.

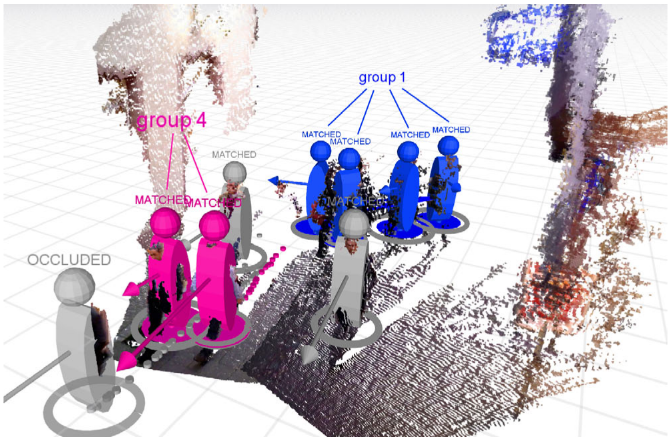

Fig. 1. Groups of four (blue) and two (pink) persons being tracked by our multi-model hypothesis tracker in RGB-D data from a crowded pedestrian area.

The paper is organized as follows: After the discussion of related work in Sec. II, we present person-level detection and tracking in Sec. III. In Sec. IV, we describe group detection, the multiple group model hypothesis tracker, and the method to determine consistent group identifiers across splits and merges. Experimental results are given in Sec. VI, and Sec. VII concludes the paper.

II. Related Work

While tracking individual people is a highly studied problem in target tracking, computer vision and robotics, the problem of tracking groups of people is relatively little-explored. However, recently, the number of related works is increasing, also due to activities in the visual surveillance and social computing communities.

We first review the literature on RGB-D based people detection and tracking from a first-person perspective. Further below, we discuss related works on group-level tracking.

Affordable RGB-D cameras have become available in the last years and are gaining popularity in many close-range sensing scenarios for intelligent cars, interactive systems and robots. As opposed to stereo cameras, they do not require expensive disparity map calculations and can be made more robust against illumination changes.

Spinello and Arras [3] propose an RGB-D person detector that probabilistically fuses the outputs of a HOG classifier (histogram of oriented gradients) on the RGB image and a HOD classifier (histogram of oriented depths) on the depth image. Luber *et al.* [4] integrate this detector into a multi-hypothesis tracking framework and combine it with a target-specific on-line detector based on RGB- and D-based appearance features.

Munaro *et al.* [5] present a person detection and tracking framework in RGB-D, where the detector uses a height map-based ROI extraction mechanism and linear SVM classification using HOG features. For tracking, they perform global nearest-neighbour data association and use a joint likelihood that also incorporates an appearance-based online term calculated from histogram features.

Recently, Jafari *et al.* [6] also use a depth-based ROI extraction mechanism, but only employ HOG in the RGB image for ROIs at distances of over 7 meters. At close-range, a normalized-depth template is evaluated on the depth image at locations where the height map shows local maxima corresponding to heads of people. For person tracking, trajectory growing is performed with a bidirectional EKF in a multi-hypothesis tracker.

For group detection and tracking, we can distinguish three lines of work:

The first one, typically carried out in the social computing community, is concerned with the understanding of social situations [7, 8]. Using interpersonal distance and relative body orientation, Groh *et al.* [7] study social situation recognition of standing people from static cameras. Similarly, Cristani *et al.* [8] address the problem of social relation recognition in conversation situations. Using interpersonal distance only, they estimate pairwise stable spatial arrangements called F-formations.

A second group of works addresses social relation recognition in still images and video. Wang *et al.* [9] extract social relations from photographs. They use the knowledge that social relations between people in photographs influence their appearance and relative image position. From the learned models, they are able to predict relationships in previously unseen images. Social relations between film actors in video are estimated by Ding *et al.* [10]. A social network graph with temporal smoothing is learned using actor occurrence patterns. The approach also allows for changes in social relations over time. Choi *et al.* [11] recognize atomic activities of individuals, interaction activities of pairs, and collective activities of groups, jointly, using an energy maximization framework.

A third line of works, most related to our context, is concerned with detecting and tracking groups from image or range data. Yu *et al.* [12] address the problem of discovery and analysis of social networks from individuals tracked in surveillance videos. A social network graph is built over time from observations of interacting individuals. Social relations between persons in overhead video data are recognized by Pellegrini *et al.* [13]. They use approximate inference on a third-order graphical model to jointly reason about correct person trajectories and group memberships. Based on learned statistical models on people's behavior in groups, they also perform group-constraint prediction of motion. Leal-Taixé *et al.* [14] model social and grouping behavior from tracked individuals in video data using a minimum-cost network flow formulation. Qin *et al.* [15] improve tracking of individuals by considering social grouping in a tracklet linking approach. Using large numbers of hypothetical partitionings of people into groups, solutions are evaluated based on the geometrical similarity of trajectories of individuals with the hypothesized group.

Lau *et al.* [16] track groups of people in 2D range data and from a mobile robot. A multi-model hypothesis tracking approach is developed to estimate the formation of tracks into groups that split and merge. Groups are collapsed into single states loosing the individual person tracks. The same tracking approach is taken by Luber *et al.* [2] but, unlike Lau *et al.*, they hypothesize and track social groupings as collections of individual person tracks with group affiliation estimates. These estimates are then used to improve person-level tracking by adapting per-target occlusion probabilities and predicting the motion of occluded group members through a constrained particle filter.

III. Person Detection and Tracking in RGB-D

A. *Multi-Sensor Person Detection in RGB-D*

For person detection in RGB-D from a first-person perspective, we use the Combo-HOD detector as presented in [3], a GPU-accelerated combination of HOG and HOD (histogram of oriented depths). HOD locally encodes the direction of depth changes and follows the same principle as HOG, while operating on the depth image. After subdividing the search window into cells, a descriptor is computed for each cell, the oriented depth gradients are collected into 1D histograms, and cells are grouped into blocks of four which then get normalized. The resulting HOD features are used to train a soft linear SVM. A depth-informed scale-space search is used to accelerate the sliding-window-based detection process.

For the sake of generality, we consider sensory setups with more than one RGB-D sensor. Thus, to fuse information from multiple sensors (such as the setups in [3, 5]), we use an individual person detector instance per sensor but combine the output of both detectors before tracking (Fig. 2). This has the advantage that the existing single-image detector can be readily used without having to fuse the RGB-D raw data which would raise the issues of

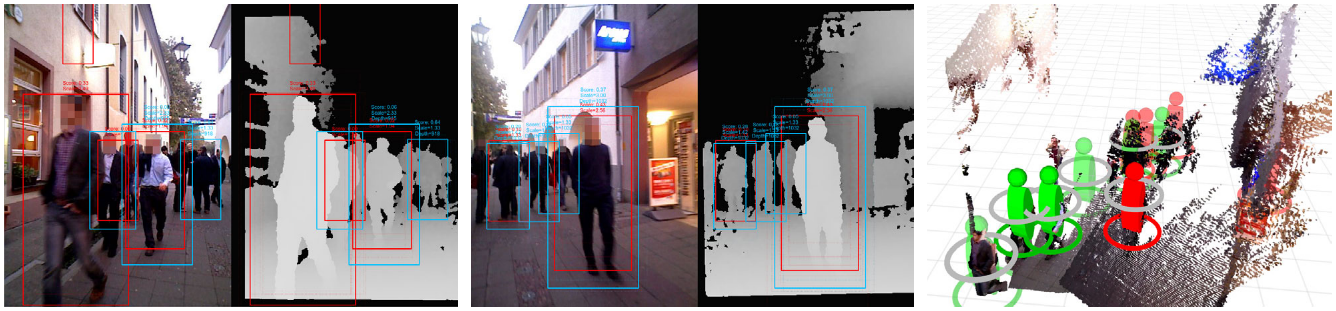

Fig. 2. *Left and middle:* RGB and depth images of left and right RGB-D sensor with HOG detections shown in red and HOD detections in blue. *Right:* Point cloud with persons detected by the left (green) and right (red) sensor. Detections where only one of the two detectors has fired are semitransparent. Overlapping detections at the common sensor boundary are merged using non-maxima suppression. Grey circles denote the resulting fused detection candidates.

image mosaicing or point cloud registration. Furthermore, a practical advantage is that each detector can be run on a single CPU core or graphics card allowing for easy parallelization. In case of a small overlap of the field of views between adjacent sensors (as in the case of our setup), we can detect people right at the center between both sensors. In a non-maxima suppression step, we finally consolidate duplicate detection hypotheses that are seen by both sensors.

B. Person Tracking using a MHT

Person detections are tracked using a multiple hypothesis tracker (MHT) due to Reid [17] and Cox *et al.* [18]. To summarize, the approach generates hypotheses about the state of the world by taking into account all statistically feasible assignments between measurements and tracks, as well as all possible interpretations of measurements as false alarms or new tracks, and of tracks as matched, occluded or deleted. At time step t, a hypothesisΩ^t_i represents one possible set of such assignments, and measurement and track interpretation labels. We call $Z(t)$ the set of detected persons at step t, $\psi_i(t)$ the predicted track-to-measurement assignments and Z^t the aggregated set of all measurements up to t. Given a parent hypothesis $\Omega^{t-1}_{l(i)}$ with index $l(i)$ to accommodate for pruning, and new incoming measurements $Z(t)$, the MHT creates new assignment sets $\psi_i(t)$, each of which gives rise to a new child hypothesis branching off from its parent. To prune the resulting exponentially growing hypotheses tree, a probability is assigned to each hypothesis that is calculated in a recursive fashion using a normalizer η, the measurement likelihood, the assignment set probability and the probability of the parent hypothesis as shown in [17]:

$$\begin{aligned} p(\Omega^t_i \mid Z^t) \quad = \quad & \eta \cdot p(Z(t) \mid \psi_i(t), \Omega^{t-1}_{l(i)}) \cdot \qquad (1) \\ & p(\psi_i(t) \mid \Omega^{t-1}_{l(i)}, Z^{t-1}) \cdot p(\Omega^{t-1}_{l(i)} \mid Z^{t-1}). \end{aligned}$$

For pruning, we use the multi-parent k-best branching according to Murty [19] and N-scan back pruning [18]. A standard Kalman filter with a constant-velocity motion model is used to predict the state of person tracks.

Particularly with a sensor in a first-person perspective, occlusions and self-occlusions of persons and group members are frequent and may last for extended periods in time. We found that the MHT with its ability to disambiguate origin uncertainty by delaying decisions until sufficient information has arrived gives good levels of robustness with respect to these events.

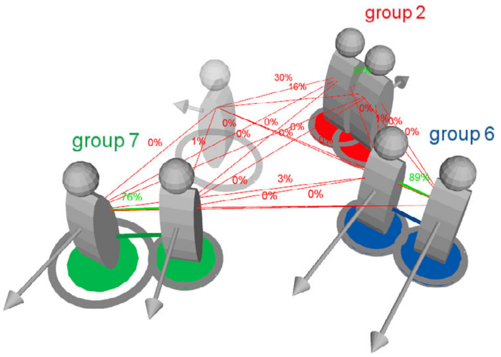

Fig. 3. A social network graph between person-level tracks. Strong social relations (indicating a group affiliation) are shown in green, weak relations in red. The relation probabilities are determined by a probabilistic SVM trained on 3D motion indicator features, and are incorporated into the group model probability (eqn. 2). The semi-transparent person is occluded but still considered during construction of the social network graph as it might reappear.

IV. Group Detection and Modeling

A. Group Detection

We detect groups by constructing and pruning a social network graph between all current targets whose edges denote pairwise social relation probabilities.

Social relation probabilities $\mathcal{R}^{i,j}$ between person track i and j, labelled as "socially related" or "not socially related" in ground truth training data, are calculated using a probabilistic support vector machine (SVM) classifier [20] trained on *coherent motion indicators*. These are motion-related features that were found to indicate group affiliation between people in large-scale empirical experiments in crowd behavior analysis and social science [1]. Concretely, coherent motion indicators consist of relative spatial distance, difference in velocity and difference in orientation of two given tracks.

All edges of the graph below a threshold of 0.5 are discarded and the remaining connected components are considered to be socially related, and thus form a group.

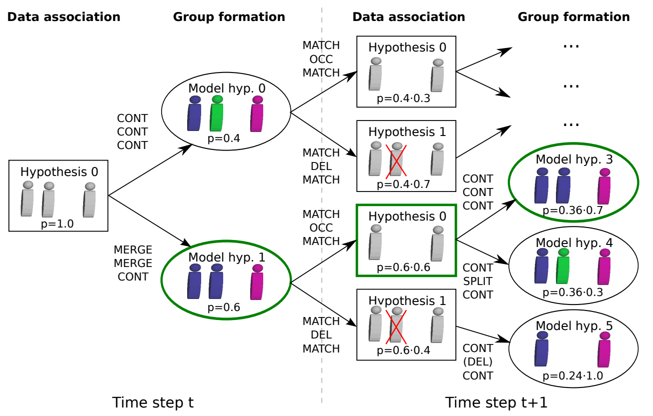

Fig. 4. An example hypothesis tree with an intermediate tree level that corresponds to a group model hypothesis step. For each of the k person-level data association hypotheses in each step, the l most probable group model hypotheses are generated that postulate different group continuation, merge and split events. The green borders indicate the maximum probability hypotheses, persons with the same color are in the same group.

The threshold was chosen because it retains pairwise social relation candidates above chance. Figure 3 shows an example social network graph from real-world data, where the edges with a strong relation (larger than the threshold) are shown in green.

Clearly, using coherent motion indicators as features to predict group affiliations implies the rather simplistic definition of "groups" as collectives of individuals that are spatially close and share a common motion goal. However, note that this group detection method scales with more available cues such as age, gender, body pose or other attributes that may indicate membership of a social group.

B. Group Model Generation

People undergo complex group formations and our goal is to track those formation processes over time. We use *merge*, *split* and *continuation* events to model the dynamic nature of group formations. The events are treated as binary operations in the sense that in a single time step, a group may split into only two groups and only two groups may merge into one group. This is a weak assumption even in the case when an entire group enters the sensor field of view at once: single-person groups will be initialized from the new tracks and after their Kalman filters have reached steady state after 4–5 cycles, the single-person groups will correctly merge into one group.

A group $\mathcal{G}_j$ becomes part of the set of *merge* candidates $\mathcal{M}_i$ of group $\mathcal{G}_i$ when the corresponding two components in the social network graph become connected (above the probability threshold of 0.5). Likewise, when two formerly connected components become disconnected, the pair of sub-groups $\mathcal{G}_i', \mathcal{G}_i''$ is added to the set of *split* candidates $\mathcal{S}_i$ belonging to group $\mathcal{G}_i$. Note that, although the chosen threshold is just above chance, this will not cause oscillations between merge and split because the social relation probabilities (edge weights) also influence the group model data likelihood term that will be discussed shortly. Finally, a group belongs to the set of *continuation* candidates $\mathcal{C}$ if it continues as is, without being involved in any split or merge event. $\mathcal{C}$ and all $\mathcal{M}_i$, $\mathcal{S}_i$ are re-initialized every cycle for each single parent data association hypothesis.

A *group model* $M(t)$ at time t represents the current group formation state, which has evolved over time through continuation, merge and split events of all groups in the scene. Formally, $M(t)$ is a partitioning of the set of all tracks at time t into groups. The multi-model hypothesis extension of the MHT now consists in a model hypothesis step interleaved with the regular observation-to-track data association step (Fig. 4). In each model hypothesis step, one or multiple group models are generated for each data association hypothesis, reflecting different possible evolutions of the parent group model $M(t-1)$ that are feasible given the social network graph at time t.

C. Group Model Probability

To generate and prune group model hypotheses, we require a model for their probability. The a priori probabilities of the group formation events *continuation* p_C, *split* p_S, and *merge* p_M, are learned from annotated real-world data sets – like [16, 2]. We assume continuation, merge and split events to be independent. Then, the probability of a given group model $M(t)$ at step t follows from the a priori probabilities of the group formation postulated by the model and from the data-driven likelihoods in the social network graph at time t. Concretely, the probability of group model $M(t)$ conditioned on its parent hypothesis Ω^{t-1} is

$$\begin{aligned} p(M(t) \mid \Omega^{t-1}) \;=\; & \underbrace{\prod_{\mathcal{G}_i \in \mathcal{C}} p_C \, p_C^{\mathcal{G}_i}}_{\text{continuations}} \cdot \\ & \underbrace{\prod_{\substack{\mathcal{G}_i \in \overline{\mathcal{C}} \\ \mathcal{G}_i', \mathcal{G}_i'' \in \mathcal{S}_i}} p_S \, p_S^{\mathcal{G}_i'\mathcal{G}_i''}}_{\text{splits}} \cdot \underbrace{\prod_{\substack{\mathcal{G}_i \in \overline{\mathcal{C}} \\ \mathcal{G}_j \in \mathcal{M}_i}} p_M \, p_M^{\mathcal{G}_i\mathcal{G}_j}}_{\text{merges}}, \end{aligned} \tag{2}$$

with the data-driven likelihoods

$$\begin{aligned} p_S^{\mathcal{G}_i'\mathcal{G}_i''} &= 1 - \mathcal{R}_{max}^{\mathcal{G}_i'\mathcal{G}_i''} \\ p_M^{\mathcal{G}_i\mathcal{G}_j} &= \mathcal{R}_{max}^{\mathcal{G}_i\mathcal{G}_j} \\ p_C^{\mathcal{G}_i} &= 1 - \max_{\substack{\mathcal{G}_j \in \mathcal{M}_i \\ \mathcal{G}_i'\mathcal{G}_i'' \in \mathcal{S}_i}} \{ p_S^{\mathcal{G}_i'\mathcal{G}_i''}, p_M^{\mathcal{G}_i\mathcal{G}_j} \} \end{aligned} \tag{3}$$

where $\overline{\mathcal{C}}$ is the complement of the set $\mathcal{C}$ of all continued groups for the current parent data association$\Omega^{\ t-1}$.

As can be seen in Eq. 3, the probability for a split of group $\mathcal{G}_i$ into sub-groups $\mathcal{G}_i'$ and $\mathcal{G}_i''$ depends on the strongest social relation between these groups, $\mathcal{R}_{max}^{\mathcal{G}_i'\mathcal{G}_i''}$. Likewise, the probability for a merge between two groups

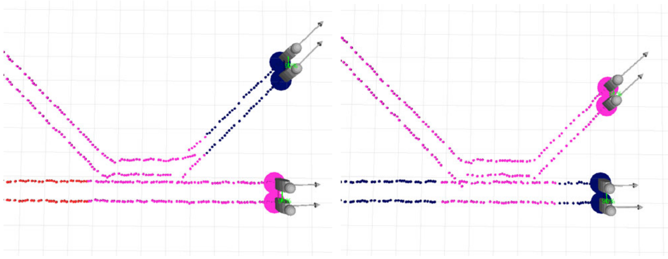

Fig. 5. *Left:* Two groups that (wrongly) merge into one group, and then split again. Without the group identifier lookup described in Sec. V-A, one of the groups undergoes an ID switch during the split (shown as different trace color). *Right:* With our extension, the previous IDs are restored.

$\mathcal{G}_i$ and $\mathcal{G}_j$ is equal to the highest probability for a person-to-person relation between these groups $\mathcal{R}_{max}^{\mathcal{G}_i\mathcal{G}_j}$. The probabilities $\mathcal{R}_{max}^{\mathcal{G}_i'\mathcal{G}_i''}$ and $\mathcal{R}_{max}^{\mathcal{G}_i\mathcal{G}_j}$ are readily available as SVM output probabilities in the social network graph.

Here, we extend the approach of [16, 2] by the data-driven probability for continuation events $p_C^{\mathcal{G}_i}$. The likelihood for continuation scales inversely with the highest probability for a non-continuation of that group (if, for instance, $p_M^{\mathcal{G}_i\mathcal{G}_j} = 1.0$, then $p_C^{\mathcal{G}_i}$ must be 0.0). Without this term, continuation events are overly biased which causes the tracker to follow splits and merges with unnecessary delays.

V. Tracking Groups of People using a Multi-Model MHT

To enhance the MHT with the ability to reason about group models, the hypothesis tree is extended by an intermediate tree level to hypothesize about possible group formation processes. Group model hypotheses spring off from data association hypotheses in each step (Fig. 4). Each group model hypothesis, in turn, will give rise to a number of child data association hypotheses conditioned on that particular group model. To limit growth of the tree, we perform multi-parent k-best branching and restrict the number of possible group models to the $l < k$ most probable models.

The incorporation of the group model probability in Eq. 2 into the recursive update rule for the probability of a hypothesisΩ_i^t can be shown to be

$$\begin{aligned} p(\Omega_i^t \mid Z^t) \quad &= \quad \eta \cdot p(Z(t) \mid \psi_i(t), M(t), \Omega_{l(i)}^{t-1}) \\ &\quad \cdot p(\psi_i(t) \mid M(t), \Omega_{l(i)}^{t-1}, Z^{t-1}) \\ &\quad \cdot p(M(t) \mid \Omega_{l(i)}^{t-1}) \cdot p(\Omega_{l(i)}^{t-1} \mid Z^{t-1}). \end{aligned} \tag{4}$$

A. Maintaining Group Identities

One of the goals of tracking is to maintain correct track identities despite misdetections, occlusions or measurement origin uncertainty. To achieve this goal in the case of group tracking, we extend [2] and define a set of maintenance rules for group track identifiers (ID) that are robust against identifier switches on the individual person track level. Group IDs need to be consolidated when groups undergo merge or split operations.

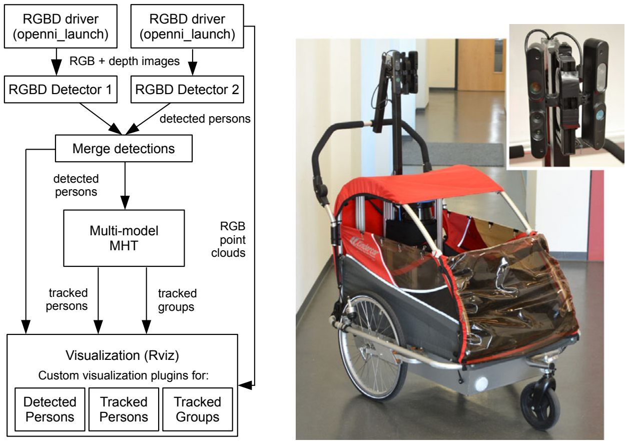

Fig. 6. *Left:* The ROS-based architecture of our tracker implementation consists of two separate RGB-D detectors, a module which merges their detection hypotheses, the multi-model MHT for person and group tracking, and several custom RViz visualization plugins. *Right:* Our mobile data capture platform equipped with wheel encoders for odometry, a laser range finder, two RGB-D sensors, and a laptop was used to capture the dataset in a urban pedestrian area. The vertical setup of two Asus Xtion Pro RGB-D sensors is shown enlarged.

Concretely, for merge events, we continue with the ID of the larger group – following a “merge-into” policy. If the groups are of the same size, we continue with the ID of the older group. This helps to maintain group ID consistency across identity switches of its member tracks, as long as not all person tracks undergo an ID switch at the same moment.

For split events, we maintain the previous social groupings in a memory including their assigned group IDs. This allows to reassign the correct ID when a group merges into another and then splits off again. The strategy is useful to reidentify sub-groups that wrongly merged with a different group, for example when groups come temporarily close in a narrow passage. The memory is implemented as a map with circular buffer, keeping the last n group ID assignments (with $n = 1000$). Fig. 5 shows the effect of our extension.

B. Group Formation Feedback into Person-Level Tracking

The information about tracked groups is fed back into the person-level tracker by adapting the per-track occlusion probabilities for group members. As shown in [2], this can effectively improve tracking performance by reducing the number identifier switches of person track, which, in turn, can make group tracking more robust. The adaptation of occlusion probabilities is done using a slight reformulation of the MHT proposed in [21], which allows the MHT to not only reason about the interpretation of tracks to be detected or deleted (as in [17], [18]) but also to be occluded. The reformulation generalizes track interpretation to an arbitrary number of labels using a multinomial distribution.

Note that the feedback of group-level information to the person-level tracker is possible because the multi-model MHT generates group formation hypotheses for each new data association hypothesis. This means that when a best hypothesis switch occurs because new evidence has made the current branch unlikely, the group tracker will instantly have the best group model available which has already evolved over time with all observations up to time t. This would not be possible if only a global, single-hypothesis group model had been generated.

VI. Experiments

For our experiments, we collected a sequence of 9 minutes, covering 300 meters of distance in an urban pedestrian zone in the city center of Freiburg, Germany. The data collection platform was a child stroller equipped with two Asus Xtion Pro Live RGB-D sensors, a SICK LMS-500 laser range finder (not used in these experiments), custom-built wheel encoder-based odometry, and a laptop (see Fig. 6). The RGB-D sensors were mounted vertically at a height of 1.4 m with about 10° of overlap, yielding a field of view of 76° in horizontal and 57° in vertically direction.

The sequences were recorded shortly before sunset to reduce IR interference, which can render the depth sensors inoperable at the cost of slightly longer exposure times. The data have been manually annotated: they contain 298 persons in 204 groups in total, of which 130 (64%) are individuals (single-person groups), 65 (32%) two-person groups, and 9 (4%) groups of at least three persons. The largest group within sensor range has 9 persons. The average group size is 1.5 – if only groups of more than two people are considered, the average size is 2.3 persons.

A. Modular ROS-based Architecture

The tracker is implemented in C++ under Linux. To facilitate use of different detectors and different sensors, clear separation of functionalities and easier parallelization, we have integrated it into a fully modular architecture using the Robot Operating System (ROS) middleware. Each module (one detector per sensor, the filter that merges detections from multiple sensors, the multi-model MHT, and the visualization) is executed in a separate process, potentially on a dedicated CPU core, and communicates with other modules via messages published on different *topics*. Fig. 6 (left) gives an overview of our architecture. The group tracker including RGB-D person detection is running on a single Intel Core i7-2600 quad-core PC at 3.4 GHz with a GeForce GTX480 graphics card.

B. Results

Our group tracking algorithm is able to track groups of people with varying sizes in crowded environments over long distances with a low number of group ID switches. Figures 7 and 8 show some exemplary situations which our tracker is able to handle. This is partly achieved due to the robustness of the person-level tracking, which benefits from the incorporated group information. Also, the probabilistic SVM based upon motion indicators allows our group detection stage to discriminate between groups passing close by each other at different velocities or in different directions. In our real-world dataset, over a duration of 9 minutes (9535 frames), we observed 11 group ID switches in total for the 74 groups of size 2 or larger.

C. Runtime Performance

The person-level MHT tracker without the RGB-D person detector reaches an average cycle time of 59 Hz on a single CPU core of a standard PC with 100 parallel hypotheses and 10 to 20 people visible at a time. This rate decreases to 24 Hz for the multi-model hypothesis group tracker, where most of the overhead is caused by the prediction of pairwise social relations using the linear SVM classifier and the update of the particle filters for the curvilinear motion model used for motion prediction of occluded group members. The memory consumption is around 2 GB. The entire system (group-level tracking and detector) reaches a cycle time of about 20 Hz.

VII. Conclusions

In this paper we addressed the problem of detecting and tracking groups of people in RGB-D data. Groups are detected from predicted social relation probabilities between individuals and tracked using an extension of the multi-hypothesis tracker that incorporates a group model hypothesis step. This approach allows to recursively reason about both, regular observation-to-track data associations and group formation processes at the same time and in the same probabilistic framework. We extended this approach by new expressions for group model probabilities and a book keeping logic to maintain stable group track identifiers robust to sporadic identifier switches of the underlying person tracks. The experiments demonstrate the viability of the approach on a real-world, unscripted outdoor RGB-D dataset collected with a mobile platform in a busy urban pedestrian zone.

In future work, we plan to deploy our group tracking framework onto a robot platform with additional backward-facing RGB-D sensors and to learn classifiers for human attributes such as age, gender and upper body pose. In addition to motion indicators, these cues will enable us to estimate social relations (and hence social groupings) more accurately, and to support data association by target-specific appearance models.

Acknowledgment

This work has been supported by the EC under contract number FP7-ICT-600877 (SPENCER).

References

[1] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, “The walking behaviour of pedestrian social groups and its impact on crowd dynamics,” *PLoS ONE*, vol. 5, no. 4, April 2010.

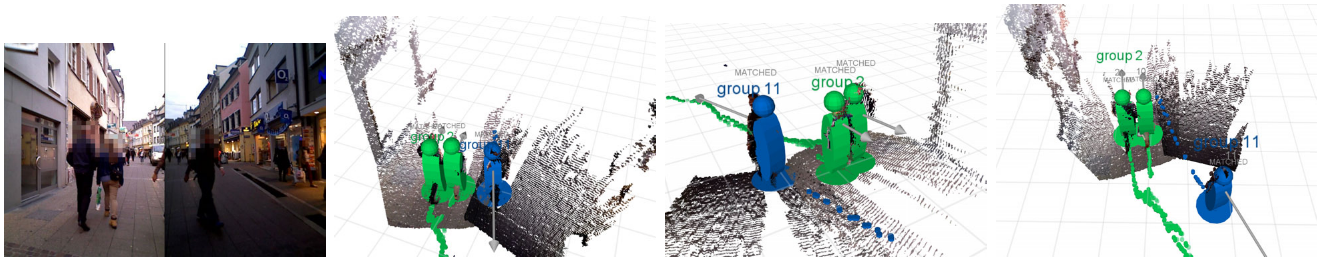

Fig. 7. Two groups – a two-person group (green) and a single-person group (blue) – are passing close-by each other, and are correctly not merged by our tracker. The SVM-based group detection using coherent motion indicators identifies them as two distinct groups of persons.

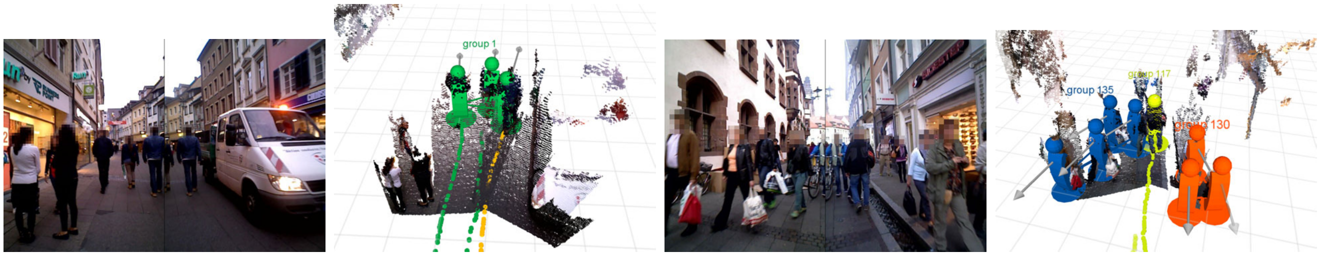

Fig. 8. *Left and middle left:* The person in blue yeans (yellow track) merges into the two-person group walking in front (green). *Middle right and right:* Our approach is able to track complex multi-person group formations in crowded environments from a first-person perspective.

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

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

[4] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGB-D data with online-boosted target models," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, San Francisco, USA, 2011.

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

[6] 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 *Int. Conf. on Robotics & Automation (ICRA)*, Hong Kong, China, 2014.

[7] G. Groh, A. Lehmann, J. Reimers, M. R. Friess, and L. Schwarz, "Detecting social situations from interaction geometry," in *Proc. of the IEEE Int. Conf. on Social Computing*, 2010.

[8] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino, "Towards computational proxemics: Inferring social relations from interpersonal distances," in *IEEE Int. Conf. on Social Computing*, Boston, USA, 2011.

[9] G. Wang, A. Gallagher, J. Luo, and D. Forsyth, "Seeing people in social context: recognizing people and social relationships," in *Proc. of the European Conf. on Comp. Vision (ECCV)*, 2010.

[10] L. Ding and A. Yilmaz, "Inferring social relations from visual concepts," in *IEEE Int. Conf. on Comp. Vis. (ICCV)*, 2011.

[11] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Proc. of the European Conf. on Comp. Vision (ECCV)*, 2012.

[12] T. Yu, S. N. Lim, K. A. Patwardhan, and N. Krahnstoever, "Monitoring, recognizing and discovering social networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[13] S. Pellegrini, A. Ess, and L. van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *Proc. of the European Conf. on Comp. Vision (ECCV)*, 2010.

[14] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker," in *ICCV Workshop on Modeling, Simul. and Vis. Analysis of Large Crowds*, 2011.

[15] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] B. Lau, K. O. Arras, and W. Burgard, "Tracking groups of people with a multi-model hypothesis tracker," in *Int. Conf. on Robotics & Automation (ICRA)*, Kobe, Japan, 2009.

[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. Pattern Anal. Mach. Intell. (PAMI)*, vol. 18, no. 2, pp. 138–150, 1996.

[19] I. Cox and M. Miller, "On finding ranked assignments with application to multi-target tracking and motion correspondence," *IEEE Trans. on Aerospace and Elect. Sys.*, vol. 31, no. 1, pp. 486–489, 1995.

[20] J. C. Platt, *Advances in Large-Margin Classifiers: Probabilities for SV Machines.* MIT Press, 2000.

[21] 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 *Int. Conf. on Robotics & Automation (ICRA)*, 2008.