Multi-Hypothesis Social Grouping and Tracking for Mobile Robots

Matthias Luber Kai O. Arras

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

***Abstract*—Detecting and tracking people and groups of people is a key skill for robots in populated environments. In this paper, we address the problem of detecting and learning socio-spatial relations between individuals and to track their group formations. 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 only 2D range data. The method, that relies on an extended multi-hypothesis tracking approach, also improves person-level tracking in two ways: the social grouping information is fed back to predict human motion over learned intra-group constraints and to support data association by adapting track-specific occlusion probabilities. Both measures lead to an improved occlusion handling and a better trade-off between false negative and false positive tracks. In experiments with a mobile robot and on large-scale outdoor data sets, we demonstrate how the approach is able to model social grouping and to improve person tracking by a significant reduction of track identifier switches and false negative tracks.**

I. Introduction

As robots enter domains in which they interact and cooperate closely with humans, people tracking becomes a key technology for many research and application areas in robotics, intelligent vehicles and interactive systems. A difficult problem is maintaining the identity of persons in crowded scenarios. Such scenarios are highly important since people typically form groups as investigated in empirical experiments where it was found that up to 70% of pedestrians walk in groups [13]. The problem of detecting, analyzing and tracking groups of people, particularly from mobile platforms, is relevant for a number of scenarios including multi-party human-robot interaction and collaboration, efficient and socially compliant robot navigation among people, analysis of social group activities, and understanding of social situations.

We first state the problem as an estimation problem of social relations between individuals from perceived track motion features using SVM classifiers and Bayesian smoothing. Since the spatial organization of groups is typically not random and remains largely stable over time, we also learn group-specific geometric relations between individuals. Opposed to prior works in which relations and social groupings are only detected on a per-frame basis or found in an a posteriori fashion by batch methods, we explicitly model and track group formations over time in a online, recursive multi-hypothesis model selection and data association framework. The recursiveness implies an anytime property where the tracker always provides a current best (but suboptimal) estimate that is refined with more incoming information. Using multi-hypothesis tracking, this happens by backtracking to branches in the hypothesis tree that become more probable with the new evidence. In contrast to batch methods, this property is crucial for mobile robots that need to take real-time decisions for interaction or navigation in unfolding social situations. The proposed approach, evaluated on 2D range data, results in accurate and fast social grouping estimates and outperforms several multi-hypothesis tracker variants.

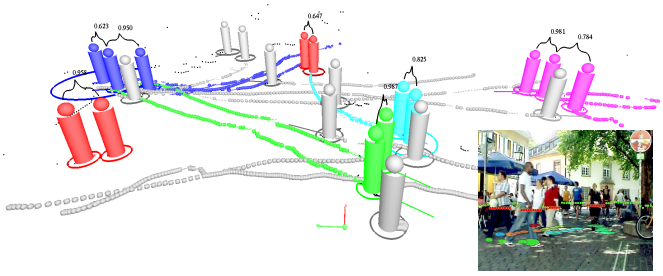

Fig. 1. A situation in the city center experiment with 23 tracks in 6 groups (shown in the same color) and several individuals (grey). Cylinders and dots denote position and trajectories of tracked persons, the numbers on top of the cylinders show social relation probabilities. Grey individuals that appear close to groups are correctly recognized to not belong to the groups as their motion properties, shown by the traces, are different. The bottom-right picture is a visualization of the scene, no image data have been used.

The paper is organized as follows: after the discussion of related work in the next section, we present the method for detecting and learning socio-spatial relations in Sec. III. Sec. IV describes how social grouping models are generated and in Sec. V we present the multi-hypothesis tracking approach. Sec. VI explains how person-level tracking is improved using the social grouping information, followed by Sec. VII that gives the experimental results.

II. Related Work

Social grouping from sensory data has recently gained increasing attention by researchers from the computer vision and social computing communities. One group of works is concerned with the understanding of social situations [9, 7]. Using interpersonal distance and relative body

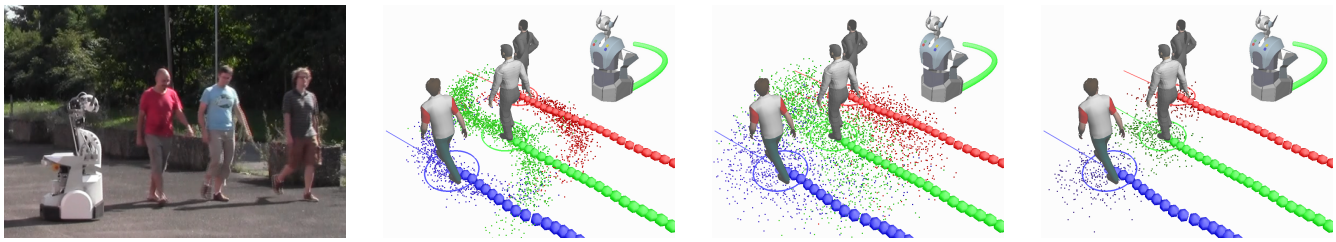

Fig. 2. On-line learning of geometric intra-group relations for three persons observed by a moving robot (*left*). The three figures on the right show initialization, prediction, and update of the spatial distributions. Empirical results from [13] for interpersonal distances and angles is used for initialization. The S-shaped distribution for the person in the center is due to the constraints from both neighbors. Spatial relations are predicted with a Brownian model and updated as soon as new tracking information is available.

orientation, Groh *et al.* [9] study social situation recognition of standing people from static cameras. Similarly, Cristani *et al.* [7] 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 addresses social relation recognition in still images and video. Wang *et al.* [19] 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.* [8]. 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.* [4] recognize atomic activities of individuals, interaction activities of pairs, and collective activities of groups, jointly, using an energy maximization framework.

A third group, most related to our context, is concerned with detecting and tracking groups from image or range data. Yu *et al.* [20] 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.* [14]. 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.* [11] model social and grouping behavior from tracked individuals in video data using a minimum-cost network flow formulation. Qin *et al.* [16] 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.* [10] 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. A very similar multi-model hypothesis approach has been developed independently by Chang *et al.* [3] to track and group neural signals whose locations are inferred from clusters of observations.

In contrast, we extend the state of the art as follows:

Opposed to [9, 7, 20, 4, 14, 11, 16] which rely on static overhead cameras to perceive the scene, we address the problem of social grouping and tracking from a mobile sensor and a first-person perspective. Overhead cameras are sufficient for surveillance but fall short of scenarios where a robot, an intelligent vehicle or an interactive system coexists and acts in the same space with people. Occlusions and misdetections occur much more often from an in-scene view than in an overhead setup. Thus, our goal is to make tracking particularly robust with respect to lengthly occlusion events and the mobility of the sensor.

Additionally, we address the problem using 2D range data which add the difficulty that targets have the same appearance and cannot be distinguished to guide data association. The use of 2D range data is relevant for robots and intelligent vehicles where such sensors are used due to their large field of view and robustness with respect to illumination and vibration. The proposed approach can obviously be applied to other sensory data, too. Furthermore, unlike [20, 14, 11, 16, 4] which employ (partly very slow) batch methods, our approach runs recursively and in real-time on a laptop PC – again highly relevant for interactive robots or vehicles that need to respond quickly to group formation changes and unpredictable events.

Unlike [9, 7] that detect and analyze social relations on a per-frame basis, we track such relations and the inferred social groupings over time. To this end, we adopt the multi-model hypothesis tracking approach developed by Lau *et al.* [10] and Chang *et al.* [3]. The method suits our problem as it allows to simultaneously hypothesize about the clustering of tracks (models) and the assignment of measurement to tracks (data association) in a consistent probabilistic framework. However, opposed to [10] who represent groups of people in a single collapsed state without spatial extension information, our approach keeps track of both the state of individual group members and the group affiliation. This allows for a much more detailed group analysis such as estimating the group's spatial extension or understanding group activities and social situations.

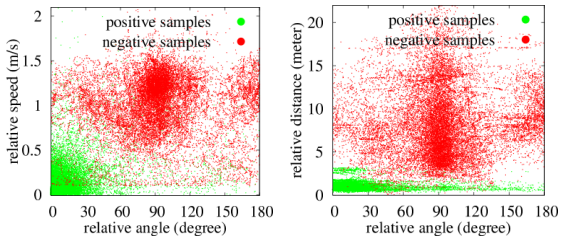

Fig. 3. Distributions of feature values of positive and negative samples in our training set. *Left:* Relative angle vs. relative speed of tracks in the same group (green dots) and tracks not in the same group (red dots). *Right:* Relative angle vs. relative distance. The distributions are consistent with the empirical findings in [12, 13].

III. Relation Learning and Group Detection

Empirical social science studies have found three dominant coherent motion indicators of people that walk in groups [12, 13]: relative distance, relative orientation, and similar velocity to their direct neighbors. Using this insight, we detect social relation candidates by classifying pairs of tracks according to their relative motion properties. The relations are used to build up a weighted social network graph in which each person corresponds to a node and edges are weighted with the probabilities of the pairwise relation. Once the graph is constructed a graph-cut algorithm is applied to extract the groups.

A. Detection of Pairwise Social Relations

A pairwise relation candidate is obtained by computing the three coherent motion indicators for two tracks of people (relative distance, angle and similar velocity) and by classifying the sample. We assume tracks of people to be represented by 2D position and velocity $\mathbf{x}_t = (x_t\, y_t\, \dot{x}_t\, \dot{y}_t)^T$ at time t with orientation $\phi_t = \text{atan2}(\dot{y}_t, \dot{x}_t)$ and velocity $v_t = \sqrt{\dot{x}_t^2 + \dot{y}_t^2}$. With $d_t^{i,j}$ being the Euclidean distance between tracks i, j, we have the feature vector $\mathcal{F}_t^{i,j} = \{d_t^{i,j}, |\phi_t^i - \phi_t^j|, |v_t^i - v_t^j|\}$ and can visualize positive and negative samples in Fig. 3. For classifying the social relation candidate $\mathcal{S}_t^{i,j}$ between i and j given $\mathcal{F}_t^{i,j}$ we use a linear SVM and map the classifier output to a probability $p(\mathcal{F}_t^{i,j}) \in [0, 1]$ according to Platt [15]. The set of social relation candidates of person i to his/her neighbors $j_1, ..., j_{N_i}$ at time t is then denoted as $\mathcal{S}_t^i = \{\mathcal{S}_t^{i,j_1}, ..., \mathcal{S}_t^{i,j_{N_i}}\}$.

This detection result relies only on information from a single frame/scan and is still noisy. Thus, we integrate the classification probabilities over time and use a simple Bayes filter to achieve smoother and more stable estimates. Let $\mathcal{F}(t) = \{\mathcal{F}_0, ..., \mathcal{F}_t\}$ be the sequence of all observed feature values of a pair of people up to the current time t, then the probability $p(\mathcal{S}_t | \mathcal{F}_t)$ of a social relation $\mathcal{S}_t$ is calculated from the parent estimate $p(\mathcal{S}_{t-1} | \mathcal{F}_{t-1})$ and the current features $\mathcal{F}_t$ in a recursive fashion using Bayes' rule $p(\mathcal{S}_t \,|\, \mathcal{F}(t)) = \eta\; p(\mathcal{F}_t \,|\, \mathcal{S}_t)\; p(\mathcal{S}_t \,|\, \mathcal{S}_{t-1})\; p(\mathcal{S}_{t-1} \,|\, \mathcal{F}(t-1))$ with $p(\mathcal{S}_t | \mathcal{S}_{t-1})$ encoding the event probabilities of social relations to arise, to be confirmed, or to break, respectively. We assume uniform priors and obtain the likelihood $p(\mathcal{F}_t | \mathcal{S}_t)$ from the one-shot detection procedure.

B. Detection of Groups

After the construction of the social network graph with filtered social relation probabilities we detect groups of people using a graph-cut algorithm. We cut the edges with probabilities lower than a threshold θ and collect all connected nodes using depth first search. The corresponding person tracks of the nodes in the subgraph are marked as group members. An example situation with six groups of two or three members is shown in Fig. 1.

C. Learning Geometric Intra-Group Relations

In addition to social relation probabilities between people, we also learn geometric intra-group relations. Moussaid *et al.* [13] investigated the spatial arrangement of pedestrians and found that people in groups form stable patterns. Thus, we can learn such patterns for person tracks in groups which amounts to estimating a track-specific spatial probability distribution of a person in the local reference frame of another person. Let $\mathbf{r}_i^j(t)$ be the geometric relation of person track i in the frame of track j and $p(\mathbf{r}_i^j(t))$ its time-dependent distribution (see Fig. 2).

We take a Monte Carlo approach to represent geometric relations since their distributions have arbitrary shapes, especially when used for human motion prediction (in Sec. VI-B). Thus, we learn the relations by recursively estimating a particle-based distribution for $p(\mathbf{r}_i^j(t) \,|\, \mathbf{z}_i^j(0), ..., \mathbf{z}_i^j(t))$ from sequences of relative track positions $\mathbf{z}_i^j(t)$.

Observations $\mathbf{z}_i^j \sim \mathcal{N}(\mu_i^j, \Sigma_i^j)$ are obtained by transforming the Gaussian state estimates from the tracker into the local frame of person i. Skipping time indices t, the mean is then computed as $\mu_i^j = \ominus H\mathbf{x}_i \oplus H\mathbf{x}_j$, a so called tail-to-tail relationship [18] with $H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$. The covariance Σ_i^j is obtained by first-order error propagation of the two state covariances Σ_i and Σ_j through the frame transform using the tail-to-tail Jacobians as derived in [18]. The proposal distribution $p(\mathbf{r}_i^j(t) | \mathbf{r}_i^j(t-1))$ is chosen to be a Brownian motion model as this model makes the least commitment for predicting the evolution of the relation. Finally, for the filter initialization priors, we take the values for $\mathbf{r}_i^j(0)$ from [13], learned from large-scale observations.

The three terms, proposal $p(\mathbf{r}_i^j(t) | \mathbf{r}_i^j(t-1))$, (sampled) Gaussian observation likelihood $p(\mathbf{z}_i^j \,|\, \mathbf{r}_i^j(t))$, and priors $p(\mathbf{r}_i^j(0))$ are then used in a particle filter with importance resampling to estimate spatial relations. A complete initialization, prediction, and update cycle in a group of three persons is illustrated in Fig. 2.

IV. Group Modeling

An important property of our approach is the ability to hypothesize about most likely graph partitions and to track them over time in a multi-model hypothesis framework. This happens recursively, based on the model of the previous step and a set of model generation events in the current step. In this section, we define those events and derive their probabilities.

Let the social network graph be $\mathcal{G}$ and the partitioning of the graph be the set of subgraphs $\{\mathcal{G}_i\}_{i=1}^{N_\mathcal{G}}$ each describing a group of people. A particular partitioning of $\mathcal{G}$ into groups is called a group model M. Each tracked person belongs to exactly one group, people without relations to others form groups of size 1.

A. Model Generation

Groups are initialized when the tracker signals a new track event, e.g. when a person enters the sensor field of view. Then, a new group of size 1 is created. Social relation computation is delayed until the track state has reached steady state in the filter, typically after four, five steps.

Once the social grouping detection is stable, each group can, at any point in time, be continued, split up into an unknown number of new groups, merged with an unknown number of other groups. Since the possible number of model transitions is large we bound this space by assuming that split and merge events are binary operations [10]. In each step, a group can only split up into two child groups and at most two groups can merge into a larger group. We also assume that a group can not be involved into a split and merge operation at the same time.

A group is terminated if all its members are declared as obsolete by the tracker. The binary operation assumption is not a sensible limitation because, for example, an instantaneous breakup of a group into three subgroups would be correctly reflected by the tracker after only two cycles.

B. Model Probabilities

The probability of a model at time t, $M(t)$, follows from the probabilities of the split, merge, and continuation events of its groups. Given the recursiveness of our problem, it conditionally depends on the model of the previous time step in parent hypothesis Ω^{t-1} at time $t-1$.

We assume constant prior probabilities for continuation events (p_C), merge events (p_M), and split events (p_S). To further narrow down the space of possible model transitions, we incorporate the actual social relation probabilities into the split and merge events, thereby implementing a data-driven aspect into the model generation step.

An existing group can only split up if none of the tracks in the two child groups share a social relation above probability threshold θ. Thus, a split event for group $\mathcal{G}_i$ occurs with a probability that scales with the strongest social relation in $\mathcal{G}_i$, $p(\mathcal{S}_{max}^{\mathcal{G}_i})$. Similarly, two groups $\mathcal{G}_i$ and $\mathcal{G}_j$ are only allowed to merge if there is at least one social relation between members of those groups greater than θ. Thus, merge events depend on the highest probability of an across-group relation, called $p(\mathcal{S}_{max}^{\mathcal{G}_i\mathcal{G}_j})$.

The conditional probability of group model $\mathcal{M}(t)$ is then

$$p(M(t) \mid \Omega^{t-1}) = p_C^{N_C} \prod_{\mathcal{G}_i} \left(p_S \left(1 - p(\mathcal{S}_{max}^{\mathcal{G}_i})\right)\right)^{\sigma_i} \prod_{\mathcal{G}_i,\mathcal{G}_j} \left(p_M \, p(\mathcal{S}_{max}^{\mathcal{G}_i\mathcal{G}_j})\right)^{\mu_{ij}}, \quad (1)$$

where N_C is the number of continued groups and σ_i and μ_{ij} are indicator variables set to 1 if $\mathcal{G}_i$ splits up or $\mathcal{G}_i, \mathcal{G}_j$ merge and 0 otherwise.

V. Multi-Model Hypothesis Tracking of Groups

In this section, we present the multi-model hypothesis tracking framework and its application to describe dynamic group formation processes.

The approach relies on the multi-hypothesis tracking approach (MHT) by Reid [17] and Cox *et al.* [5] and its extension to incorporate a multi model-hypothesis step by Lau *et al.* [10] and Chang *et al.* [3]. For reasons of limited space we give a brief summary of the method, for more details the reader is referred to the original papers.

The MHT algorithm hypothesizes about the state of the world 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. A hypothesis Ω_i^t is one possible set of assignments and interpretations at time t. Let $Z(t) = \{z_i(t)\}_{i=1}^{m_t}$ be the set of m_t detected people and $\psi_i(t)$ the set of assignments which associates predicted tracks to measurements in $Z(t)$. Further, let Z^t be the set of all measurements up to time t. Starting from a hypothesis of the previous time step, called a parent hypothesis $\Omega_{l(i)}^{t-1}$, and a new set $Z(t)$, there are many possible assignment sets $\psi(t)$, each giving birth to a child hypothesis that branches off the parent. This makes up an exponentially growing hypothesis tree. For a real-time implementation, the growing tree needs to be pruned. To guide the pruning, each hypothesis receives a probability, recursively calculated as the product of a normalizer η, a measurement likelihood, an assignment set probability and the parent hypothesis probability,

$$p(\Omega_i^t \mid Z^t) = \eta \, p(Z(t) \mid \psi_i(t), \Omega_{l(i)}^{t-1}) \; p(\psi_i(t) \mid \Omega_{l(i)}^{t-1}, Z^{t-1}) \; p(\Omega_{l(i)}^{t-1} \mid Z^{t-1}). \quad (2)$$

The extension to a multi-model tracking approach that hypothesizes over both, data associations and models is as follows: the multi-model MHT introduces an *intermediate tree level at each time step*, on which models spring off from parent hypotheses. Each model branch has its own data association tree, conditioned on that model. Formally, this adds a model probability term to Eq. 2 and introduces the model as a conditioning variable (see derivation in [10]),

$$p(\Omega_i^t \mid Z^t) = \eta \, p(Z(t) \mid \psi_i(t), M(t), \Omega_{l(i)}^{t-1}) \; p(\psi_i(t) \mid M(t), \Omega_{l(i)}^{t-1}, Z^{t-1}) \; p(M(t) \mid \Omega_{l(i)}^{t-1}) \; p(\Omega_{l(i)}^{t-1} \mid Z^{t-1}). \quad (3)$$

The model probability term is defined in Eq. 1. The final expression for Eq. 3 is rather simple as many variables cancel out after substitution. This property is retained with the multi-model hypothesis extension.

For pruning this tree, we pursue two effective strategies: on the level of the regular data association MHT, we

employ multi-parent k-best branching according to [6] which generates only the global k most probable hypotheses in polynomial time. Furthermore, in our experiments we have found that parents may branch into multiple child models of similar (high) probability, causing the tracker to loose diversity since the tree concentrates on few probable branches. Thus, we also bound model branching to the most $l \leq k$ most probable models that arise from a common parent hypothesis.

VI. Group-Enabled People Tracking

At this point, we are able to find and keep track of groups of people and hypothesize about their formation processes. This knowledge is already relevant for human-robot interaction or robot navigation tasks among people. In addition to this, we can feed back the social grouping information to improve tracking on the person level: we use the social relation information to adapt the occlusion probabilities of individual tracks in groups and make constraint motion predictions for such tracks via the learned spatial intra-group relations.

A. Integration of Social Relations

When perceiving the scene from a first-person perspective, occlusions occur particularly often for people in groups. Thus, person tracks for which the system predicts a high social relation probability will have a higher occlusion probability. Formally, this can be implemented using a simple extension of the MHT initially developed for leg tracks in [1]. The extension allows the MHT to not only reason about the interpretation of tracks to be detected or deleted (as in [17], [5]) but also to be occluded. This implies a generalization to an arbitrary number of track interpretation labels and the modeling of their numbers in an assignment set by a multinomial distribution. With occlusion being a label on its own, we can adapt the occlusion probability of individual tracks dynamically.

This involves learning a set of constant probabilities for the final expression of $p(\Omega_i^t \,|\, Z^t)$. While the non-adaptive MHT has parameters for track detection, occlusion and deletion events, denoted p_{det}, p_{occ} and p_{del}, the MHT with adaptive occlusion probabilities requires to learn those probabilities for tracks in groups separately, $p_{det|G}$, $p_{occ|G}$ and $p_{del|G}$. Both sets, learned from large-scale datasets, are subject to the multinomial constraint $p_{det}+p_{del}+p_{occ} = 1$.

Finally, during tracking, the parameter set p_{det}, p_{del}, p_{occ} is used for tracks without social relations and the set $p_{det|G}$, $p_{del|G}$, $p_{occ|G}$ is used for person tracks in groups.

B. Integration of Geometric Relations

Geometric relations are used for better prediction of human motion in groups. Based on the observation that people in groups largely maintain their spatial organization [12, 13], the learned relations will enable us to predict occluded group members over the visible group members.

Let $\mathcal{R}_i(t) = \{\mathbf{r}_i^j(t)\}_{j=1}^{N_r}$ be the set of geometric relations of track i to the N_r neighboring tracks in the same group,

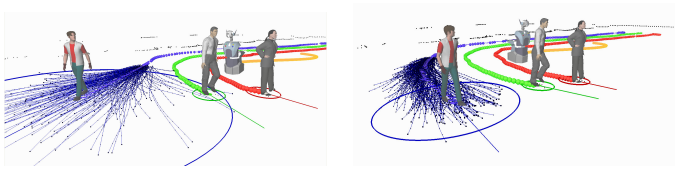

Fig. 4. Effect of the constrained motion prediction from intra-group relations. *Left:* without geometric relations, *Right:* with relations. Colored circles, dots, and lines show position uncertainties, motion particles, and trajectories, respectively. The curvilinear motion model allows to readily incorporate the on-line learned relations and maintains the spatial organization of groups also during maneuvers.

then the motion model $p(\mathbf{x}_t|\,\mathbf{x}_{t-1}, \mathcal{R}(t-1))$ becomes conditioned on both, the previous track state $\mathbf{x}_{t-1}$ and the set $\mathcal{R}(t-1)$ at time index $t-1$. According to our particle-based representation of geometric relations, we represent our target distribution with a set of weighted samples

$$p(\mathbf{x}_t|\,\mathbf{x}_{t-1}, \mathcal{R}(t-1)) \simeq \sum_i w_t^{(i)} \delta_{\mathbf{x}_t^{(i)}}(\mathbf{x}_t) \qquad (4)$$

where $\delta_{\mathbf{x}_t^{(i)}}(\mathbf{x}_t)$ is the impulse function centered at $\mathbf{x}_t^{(i)}$.

Sampling directly from this distribution is intractable in practice which is why we take a Monte Carlo approach, in which samples are first drawn from a proposal distribution π and then evaluated according to the mismatch between the target distribution τ and the proposal distribution. In our case, the distribution is approximated by the following factorization

$$p(\mathbf{x}_t|\,\mathbf{x}_{t-1}, \mathcal{R}(t-1)) \;\simeq\; p(\mathbf{x}_t|\,\mathbf{x}_{t-1})\; p(\mathbf{x}_t|\,\mathcal{R}(t-1)) \qquad (5)$$

where we choose a motion model $p(\mathbf{x}_t^{(i)}|\,\mathbf{x}_{t-1})$ as proposal distribution and evaluate the samples according to

$$w_t^{(i)} = \frac{p(\mathbf{x}_t|\,\mathbf{x}_{t-1}, \mathcal{R}(t-1))}{p(\mathbf{x}_t^{(i)}|\,\mathbf{x}_{t-1})} = p(\mathbf{x}_t^{(i)}|\,\mathcal{R}(t-1)). \qquad (6)$$

In other words, samples are first spread out into the state space by $p(\mathbf{x}_t^{(i)}|\,\mathbf{x}_{t-1})$ and then weighted according to the set of geometric relations $\mathcal{R}(t-1)$.

For $p(\mathbf{x}_t^{(i)}|\,\mathbf{x}_{t-1})$, we take the curvilinear model by Best and Norton [2] which is especially well suited for maneuvering targets. Unlike the constant velocity model it accounts for both, (cross-track) normal and (along-track) tangential target accelerations needed to properly maintain the spatial intra-group relations during direction changes of the entire group (Fig. 4). Let $\mathbf{x}_t^{(i)} = (\,x_t\;\; y_t\;\; \dot{x}_t\;\; \dot{y}_t\,)^T$ be the state of particle i, $\mathbf{a}_t = (\,a_t\;\; a_n\,)^T$ the vector of tangential and normal accelerations, and A the constant velocity transition matrix, then the particle states evolve according to

$$\mathbf{x}_{t+1}^{(i)} \;=\; A\,\mathbf{x}_t^{(i)} + G_t(\mathbf{a}_t^{(i)} + q_t) \qquad (7)$$

with zero-mean Gaussian noise q_t. The details on the 4×2 forcing matrix G_t can be found in [2].

To evaluate the likelihood $p(\mathbf{x}_t^{(i)}|\,\mathcal{R}(t-1))$ of particle i we consider the N_r geometric relations of $\mathcal{R}(t-1)$ as components of a mixture with equal mixture weights.

TABLE I

RESULTS OF THE SOCIAL RELATION DETECTION ON THE TRAINING SET (TRAINING), THE TEST SET WITHOUT BAYESIAN SMOOTHING (ONE-SHOT), AND WITH THE BAYES FILTER (FILTERED).

Approach	TP	FP	TN	FN	PR	RE	ACC
training	6342	879	12703	520	0.88	0.92	0.93
one-shot	4598	733	12849	2264	0.86	0.67	0.85
filtered	5701	1256	12326	1161	0.81	0.83	0.88

During tracking, visible group tracks are predicted using a constant velocity motion model and, after the Kalman update, are used to predict tracks of the same group that have been declared as occluded by the MHT. The occluded tracks are predicted by spreading their particles according to Eq. 7. To evaluate their weights, the geometric relations need to be transformed into the reference frame of the tracker. If all group members were occluded, motion prediction would fall back onto the constant velocity model.

VII. EXPERIMENTS

We will now evaluate the proposed tracker and analyze the contribution of all extensions to the tracking performance. The experiments are carried out on two exemplary data sets collected with our mobile robot DARYL and two large, unscripted outdoor data sets collected in a city center and a main station environment during a regular work day. The sensor is always a SICK LMS 291 laser range finder at around 0.80 m height and 0.5 deg angular resolution. The large outdoor data sets of 55,475 and 33,204 frames (25 and 15 min, respectively), recorded at fairly busy city locations, contain data on individuals, couples, groups of people, bicycles, cars, wheelchairs, skaters and person-shaped static obstacles that all undergo countless occlusions (see Fig. 1 for an example frame). The data have been manually annotated to determine the detection, data association, and social grouping ground truth. Criteria for social grouping annotations were people's trajectories, behaviors, and appearances (camera data were available). In detail, the city center data set consists of 10,000 frames with 190 persons including 31 groups. The main station set contains 6,000 frames, 168 person tracks, and 25 groups.

The person detector for 2D range data, which is a boosted feature-based classifier [1], the linear SVM for detecting social relations, and the MHT parameters have all been learned on a separate training set with 95 tracks over 28,242 frames. The MHT parameters are as follows: the detection, occlusion, and deletion probabilities are $p_{det} = 0.7$, $p_{occ} = 0.27$, and $p_{del} = 0.03$, respectively. The Poisson rates for false alarms and new tracks in the MHT are $\lambda_{new} = 0.0003$ and $\lambda_{fal} = 0.005$, and the parameters for people in groups are $p_{det|G} = 0.6$, $p_{occ|G} = 0.39$, $p_{del|G} = 0.01$. Geometric relations are learned and predicted using 200 particles per track. The maximum number of MHT hypothesis is $k = 100$ and the maximum number of model branches per hypothesis is $l = 10$.

As tracking performance measure we employ the MOTA metrics and count three numbers with respect to ground truth: misses (missing person tracks that should exist at a ground truth position, FN), false positives (person tracks that should not exist, FP), and mismatches (track identifier switches, ID). From these numbers the tracking accuracy MOTA is determined as $1 - \frac{\#err}{\#evt}$, with $\#err$ being the number of tracking errors $\#err = FN + FP + ID$ and $\#evt$ the number of tracking events over the length of the experiment. Note that due to the normalization by the total number of events, even large reductions of the errors may result in only small changes of the MOTA score.

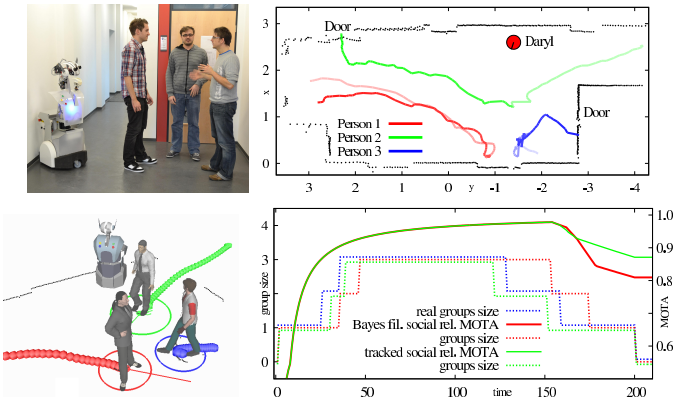

Fig. 5. Comparison of the filtered per-frame approach vs. the tracking approach. Three people meet, interact, and split up (left column) with trajectories shown in the top right image. The bottom right graph shows the group formation process and person-level tracking accuracy. The smoothing approach (red) overly delays group merge and split events and cause the accuracy (solid) to drop. Tracked social relations (green) are clearly closer to ground truth (blue).

A. Detecting Social Relations

We first evaluate the accuracy of the social relation detection and the impact of the Bayesian filtering.

On the training set, the detection accuracy of the SVM classifier is 93% with only 879 false positives (FP) and 520 false negatives (FN). During tracking on the test set, this decreases to 85% accuracy mostly due to missed social relations (2264 FN) during the track initialization phase when the orientation and velocity state estimates are not yet in steady state. Bayes filtering the social relation probabilities improves the number of misses by 50% but comes at the expense of delayed responses, e.g. when people leave a group. This causes the number of false positives to increase by almost a factor of two. The overall detection accuracy is 88%. See Tab. I for all numbers including precision (PR), recall (RE), and accuracy (ACC).

This accuracy is sufficient for our purposes. However, inferring relations only from motion features is clearly limited and future work will focus on recognizing more attributes of people as cues for social relations.

B. Tracking Social Relations

When evaluating the impact of tracking the social grouping hypotheses, we find that the proposed approach is able to resolve the trade off between lower numbers of false negatives and delayed response times. This is demonstrated in the indoor experiment in Fig. 5 where

TABLE II
TRACKING PERFORMANCE RESULTS ON THE LARGE-SCALE DATA SETS USING 100 HYPOTHESES.

Data Set	Approach	FN	FP	ID	MOTA	*Hz*
city center	baseline	4746	2344	261	79.6%	52.5
	geometric rel.	4179 (-11.9%)	2287 (-2.4%)	206 (-21.1%)	81.5%	25.2
	Bayes filtered social rel.	3407 (-28.2%)	2752 (+17.4%)	196 (-24.9%)	82.4%	27.8
	+ geometric rel.	3169 (-33.2%)	2808 (+19.8%)	171 (-34.4%)	83.0%	20.5
	tracked social rel.	3600 (-24.1%)	2387 (+1.8%)	179 (-31.4%)	82.9%	29.6
	+ geometric rel.	3472 (-26.8%)	2390 (+1.9%)	162 (-37.9%)	83.3%	17.6
main station	baseline	2949	2982	360	81.4%	31.1
	geometric rel.	2342 (-20.6%)	3138 (+5.2%)	284 (-21.1%)	83.0%	23.7
	Bayes filtered social rel.	2067 (-29.9%)	3488 (+16.9%)	245 (-31.9%)	82.9%	25.3
	+ geometric rel.	1878 (-36.3%)	3459 (+15.9%)	231 (-35.8%)	83.6%	18.8
	tracked social rel.	2122 (-28.0%)	3158 (+5.9%)	202 (-43.9%)	83.8%	23.7
	+ geometric rel.	2108 (-28.5%)	3150 (+5.6%)	193 (-46.4%)	83.9%	18.4

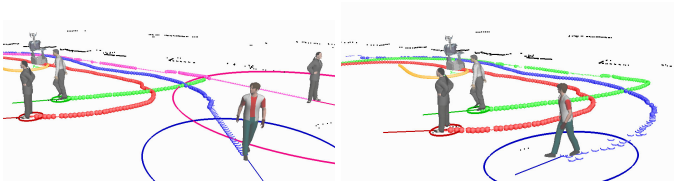

Fig. 6. The mobile robot observes a group of three persons that undergo a turn maneuver. Colored circles and dots show position uncertainties and trajectories, flat semi-circles denote occlusions. *Left:* The baseline tracker does a poor job of predicting occluded tracks during the maneuver and looses one person (colored in pink, re-initialized in green). *Right:* Using social and geometric information during tracking, the inter-group organization is maintained over occlusions and no track loss occurs.

three people meet, form a group during interaction and split up again. The multi-model hypothesis tracker is able to reflect those group formation changes much faster than the Bayesian smoothing approach (see dotted lines in bottom-right diagram). This is due to the ability of the multi-hypothesis approach to consider multiple model explanations at a time and backtrack to branches that have become more probable with more incoming information. The delayed responses of the Bayesian approach make that the group merge phase lags behind and that the persons are kept in one group overly long after the split up. The solid lines in the same diagram show the person-level tracking accuracy (MOTA) which is consistently high for the multi-hypothesis tracking approach versus a drop from 87% to 81% for the Bayesian filtering approach.

While in this experiment the improvement seems not dramatic, the faster response times are crucial for robots that reactively navigate among people.

C. Geometric and Social Relations During Tracking

In this experiment, we evaluate the adaptive occlusion probabilities, the on-line estimated geometric relations, and their ability to predict group tracks over lengthy occlusions. The baseline is the tracker without social and geometric information. During a sequence of six minutes three persons were instructed to walk and stand in the vicinity of the robot while changing their spatial arrangement. The robot was moving during the experiment.

An example situation is shown in Fig. 6 during a turn of the group. The baseline approach is unable to maintain the spatial organization of the group and looses track of one person. The total number of track losses during the experiment is 12. The proposed system with social and geometric relations maintains the spatial organization and has no track loss. The particle filter is also able to quickly adapt the on-line learned geometric relations to changes of the spatial group arrangement.

D. People Tracking using Social and Spatial Information

Finally, we evaluate the multi-model hypothesis MHT on two large-scale outdoor data sets (city center and main station). We compare it to a regular MHT without the group model hypothesis extension and study the contributions of the social and geometric relation information separately (see Tab. II). Hereafter, we discuss the results:

Geometric Relations (2nd row): Using on-line estimated geometric intra-group relations for motion prediction of tracks has a positive effect on the number of false negatives FN (-11,9%/-20,6%) and track ID switches (-21,1%), and a neutral effect on the number of false positives FP (-2,4%/+5,2%). As shown in Fig. 6, the relations allow the tracker to properly track persons during occlusions and group maneuvers. It is noteworthy that the approach finds a good trade-off between occlusion handling and an increase of false positive tracks. Naive methods handle occlusions simply by delaying the deletion of tracks but cause the number of wrong tracks (e.g. from false positive people detections) to persist longer in the system as well. Ergo, the result suggests that the approach of learning spatial group arrangements is a well suited method to deal with occlusions in our context.

Smoothed vs. Tracked Social Relations (3rd and 5th row): The Bayesian filtering approach appears to have this very problem. It improves the FN and ID measures but causes a significant increase in the FP measure. The method is too simple to find a good track management trade-off, one reason being the slow response to group formation changes as already shown in Fig. 5. In contrast, the multi-model hypothesis approach finds the so far best FP versus FN/ID trade-off. This is mainly due to the faster response times

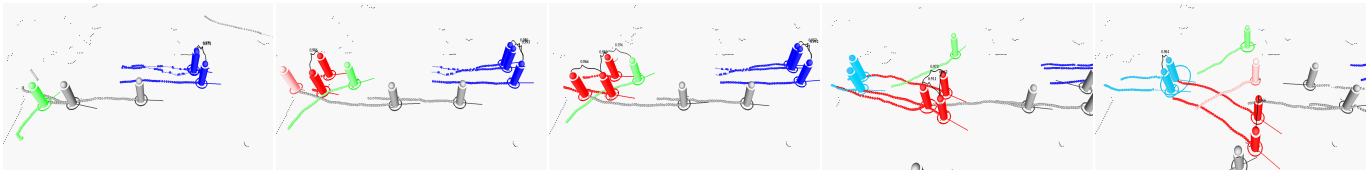

Fig. 7. Tracking sequence in the city center data set. The cylinders and dots show positions and trajectories. People in the same groups have identical color, those in grey and green share no social relations. The green-colored track (correctly) does not merge with the red group (max. social relation probability 0.396 in frame 3). The person in light red merges with the red group in frame 2 and splits up in frame 5.

to group formation changes and a proper incorporation of domain knowledge into occlusion handling.

Combining Social and Geometric Information (4th and 6th row): Adding the constraint-based motion prediction model improves both approaches (Bayes filter and multi-model hypothesis MHT) in most measures. In this setting, the multi-model hypothesis MHT yields the overall best results, expressed also by the highest MOTA scores. Notice that we achieved a key improvement over the baseline of -38%/-46% fewer track ID switches, the most important performance measure in scenarios that involve interaction with and motion prediction of people.

Finally, with a cycle time of at least $17.6\,Hz$ on a standard laptop PC, the approach is well applicable under real-time conditions. An example sequence of group formations with the tracking results in shown in Fig. 7.

VIII. Conclusions

In this paper we have addressed the problem of detecting and learning socio-spatial relations between people as well as inferring and tracking their social groupings. The proposed approach, that relies on an extension of a multi-hypothesis tracking approach, also improves person-level tracking in two ways: the social grouping information is used to predict human motion over learned intra-group constraints and to support data association by adapting track-specific occlusion probabilities.

Opposed to most related works that use static overhead cameras and batch approaches, we address the problem from a mobile platform, learn geometric relations in an on-line fashion and track group affiliations with a recursive multi-model hypothesis tracker in real-time. With roughly 40% fewer track identity switches and 28% fewer false negative tracks, the results suggest that tracking people in 2D range data can strongly benefit from estimates on social and geometrical relations, mostly due to their ability to explain lengthy occlusion events.

In future work we plan to use RGB-D data and incorporate additional attribute information on people such as age and gender to further improve social relation estimation.

Acknowledgment

This work has partly been supported by the German Research Foundation (DFG) under contract number SFB/TR-8 and by the EC under contract number FP7-ICT-600877 (SPENCER).

References

[1] 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. of the Int. Conf. on Robotics & Automation (ICRA)*, 2008.

[2] R.A. Best and J.P. Norton. A new model and efficient tracker for a target with curvilinear motion. *IEEE Transactions on Aerospace and Electronic Systems*, 33(3):1030–1037, 1997.

[3] S. Chang, R. Sharan, M. T. Wolf, N. Mitsumoto, and J. W. Burdick. People tracking with UWB radar using a multiple-hypothesis tracking of clusters (MHTC) method. *Int. Journal of Social Robotics*, 2(1):3–18, 2010.

[4] 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.

[5] 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)*, 18(2):138–150, 1996.

[6] I.J. Cox and M.L. Miller. On finding ranked assignments with application to multi-target tracking and motion correspondence. *IEEE Trans. on Aerospace and Elect. Sys.*, 31(1):486–489, 1995.

[7] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *Proc. of the 2010 IEEE Int. Conf. on Social Computing*, Boston, MA, USA, 2011.

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

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

[10] B. Lau, K. O. Arras, and W. Burgard. Multi-model hypothesis group tracking and group size estimation. *Int. Journal of Social Robotics*, 2(1):19–30, March 2010.

[11] 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.

[12] C. McPhail and R. Wohlstein. Using film to analyze pedestrian behavior. *Sociological Methods & Research*, 10(3):347–375, 1982.

[13] 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*, 5(4):e10047, April 2010.

[14] 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.

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

[16] 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.

[17] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6), 1979.

[18] R. Smith, M. Self, and P. Cheeseman. Estimating uncertain spatial realtionships in robotics. In *Autonomous Robot Vehicles*. Springer Verlag, 1990.

[19] 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.

[20] 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.