Joint Individual-Group Modeling for Tracking

Loris Bazzani, Matteo Zanotto, Marco Cristani, *Member, IEEE,*
and Vittorio Murino, *Senior Member, IEEE*

Abstract—We present a novel probabilistic framework that jointly models individuals and groups for tracking. Managing groups is challenging, primarily because of their nonlinear dynamics and complex layout which lead to repeated splitting and merging events. The proposed approach assumes a tight relation of mutual support between the modeling of individuals and groups, promoting the idea that groups are better modeled if individuals are considered and vice versa. This concept is translated in a mathematical model using a decentralized particle filtering framework which deals with a joint individual-group state space. The model factorizes the joint space into two dependent subspaces, where individuals and groups share the knowledge of the joint individual-group distribution. The assignment of people to the different groups (and thus group initialization, split and merge) is implemented by two alternative strategies: using classifiers trained beforehand on statistics of group configurations, and through online learning of a Dirichlet process mixture model, assuming that no training data is available before tracking. These strategies lead to two different methods that can be used on top of any person detector (simulated using the ground truth in our experiments). We provide convincing results on two recent challenging tracking benchmarks.

Index Terms—Group modeling, joint individual-group tracking, decentralized particle filtering, Dirichlet process mixture model

1 Introduction

People tracking is one of the most important topics in computer vision [1], [2], [3], [4], [5], representing a core module of general video analytics systems, lying between low-level processing (like background subtraction [6], object detection [7], person re-identification [8]) and high-level processing (action/activity analysis [9], social signal processing [10]).

Despite being an open problem, tracking has been recently reformulated to tackle new challenges, like crowd analysis and group tracking. Crowd analysis [11], [12], [13] is typically addressed using ad-hoc strategies which do not focus on the single individual, but on dense masses, exploiting optical flow-based strategies [13]. However, groups are relatively unlikely to stay stable over time in such scenarios where individuals follow a less predictable trajectory because of the physical constraints of the crowd.

Group tracking is instead related to a smaller number of individuals. A group is defined in [14] as *a social unit comprising several members who stand in status and relationships with one another*. There are many types of group, that differ for durability (ad hoc or stable groups), informality of organization, and level of physical dispersion [15]. In this paper, we focus on *self-organizing* groups, defined as individuals that gradually cooperate and engage with each other around some task of interest [16]. Examples of self-organizing groups are two friends that meet and discuss in an open plaza, a team of supporters going to watch their preferred team, a family exiting from the mall and going to their car, a couple visiting a museum. In practice, a self-organizing group is the result of a focused interaction, where the involved subjects spontaneously decide to be in each other's immediate presence, adopting a proxemics behavior functional to this end, that is, modulating their velocity and revising their location [17].

Tracking self-organizing groups is beneficial in many scenarios. In video surveillance, automatically understanding the network of social relationships observed in an ecological scenario might help for advanced suspect profiling. For example, understanding if a person walks in a group helps to re-identify it [18]. In social robotics, a robot can be programmed with the aim of detecting and tracking the biggest group in order to maximize the possibility of interaction [19].

Individual-group tracking is a challenging problem because groups are subject to frequent events that substantially modify their configuration. In particular, merging and splitting phenomena, absent in the individual tracking, play a crucial role. From a sociological perspective, whenever an individual leaves a group (an example of split) or joins a group (merge), all the social relationships between the remaining people are revised so that the individuals produce new entities [20]. One of the most challenging problems in individual-group tracking is to handle those merging and splitting events, since they are responsible for the creation or the deletion of groups. The proposed model is able to cope with the change of identities when groups split or merge.

In the literature, groups are modeled following two orthogonal viewpoints: in one case, groups are atomic entities, so that standard multi-person tracking methods

- *L. Bazzani, M. Zanotto, M. Cristani and V. Murino are with the Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy. E-mail: name.surname@iit.it*
- *M. Cristani and V. Murino are also with the Department of Computer Science, University of Verona, Verona, Italy.*
- *L. Bazzani and M. Zanotto contributed equally to this work.*

can be applied [21], [22], [23]. The other case considers individuals as distinct objects to track, and their tracklets are employed to find common behaviors, giving rise to group entities [24], [25], [26], [27]. Under both these perspectives, group tracking is strongly dependent on the performance of the individual tracker.

This paper proposes a method to track both groups and individuals in a *unified* framework which can be used as a module of an end-to-end vision pipeline for group behavior analysis. Specifically, the presented joint framework considers the dual nature of groups: they are entities in their own rights, but they are by definition a collection of simpler entities. This goes in the opposite direction with respect to those strategies which propose an ex-post refinement of the group hypotheses using individual statistics [25], [26], [28], or revising the posterior on the individual whereabouts with grouping dynamics [24], [29]. Our proposal instead exploits directly the dependence between the individual and the group statistics in a similar spirit of what has been done in other contexts [30], [31]. In the proposed model, the support of the individual statistics feeds the reasoning module at a group level, which itself gives feedback to the individual tracking. This allows to obtain a trade-off between individual and group tracking performance. Technically, this is made possible since the group and individual tracking instances combine their estimates in a single, joint state space.

The proposed method is built upon the recent Decentralized Particle Filter (DPF) [32] as a joint tracking framework. The DPF is an online inference method that deals with arbitrary state spaces by decomposing them in dependent subspaces. It solves the filtering problem of the separate subspaces in a nested way and transfers the uncertainty across them through a set of conditional probability distributions. We build the proposed model on this framework, allowing to divide the joint group-individuals state space in the subspace of individuals and the subspace of groups and at the same time, accounting for their mutual dependence through conditional probability distributions.

In particular, our new formulation considers two different strategies: the first one embeds in the different distributions a set of offline-trained classifiers for tracking group entities [33]. The second one fuses the DPF with a Dirichlet Process Mixture Model (DPMM) [34], that learns in an online fashion the group configuration without the need for any pre-trained classifier. This differentiation gives rise to two different Joint Individual-Group Trackers (JIGT), called DEcentralizEd Particle filtER (DEEPER-JIGT) and Dirichlet Process-based Decentralized Particle filter (DP2-JIGT).

Both approaches allow to tackle individual-group tracking in a novel way and naturally deal with splitting or merging of groups. The cooperation of these heterogeneous points of view is in general beneficial as compared to other group modeling and tracking models (*e.g.*, [24], [26]). The two strategies have been evaluated on different benchmarks, both based on synthetic and real data. We found that the DP2-JIGT performs best when groups stay in the scene long enough to allow the online learning algorithm to infer the group configurations. The DEEPER-JIGT is a valid alternative for shorter sequences. Despite this difference, both approaches show interesting characteristics when dealing with splitting or merging of groups. To evaluate them, recent metrics of group detection are employed, together with standard figures of merit adapted from the multi-person tracking literature.

The rest of the paper is organized as follows. In Sec. 2, the state of the art of individual and group modeling for tracking is analyzed. Sec. 3 formulates the problem and introduces the proposed model. Sec. 4 presents the mathematical framework for joint tracking. We describe the group models and the methods for joint individual-group tracking in Sec. 5 and Sec. 6, respectively. Experimental results are reported in Sec. 7, and, finally, Sec. 8 concludes the paper and sketches the future perspectives of the work.

2 State of the Art

The literature of group analysis is divided in two classes: one is focused on approaches that detect groups without taking into account temporal information, whereas the other considers proper group tracking techniques.

Concerning the first class, a geometrical approach which defines groups as a set of adjacent Voronoi polygons on the position of the individuals is proposed in [35]. Positional information, though, is not informative enough to capture all the characteristics of groups, and additional features are usually considered. As an example, in semi-stationary scenarios (*e.g.*, cocktail parties), head or body orientation are used to check pairwise interactions between individuals [36], [37] or to detect F-formations [38], [39]. Head pose estimation algorithms, though, are prone to errors mainly due to the low resolution of videos, background clutter and occlusions.

Tracking techniques model the temporal evolution of groups: the recent literature on group tracking can be partitioned in three categories: 1) the class of *group-based* techniques, where groups are treated as atomic entities without the support of individual tracks statistics ([21], [22], [23], [40], [41]); 2) the class of *individual-based* methods, where group descriptions are built by associating individuals' tracklets that have been calculated beforehand, typically, with a time lag of few seconds ([24], [25], [26], [27], [28], [29], [42], [43], [44]); and 3) the class of *joint individual-group* approaches, where group tracking and individual tracking are performed simultaneously ([27], [33], [45], [46], [47]). Our approaches lie in the last category, presenting characteristics of substantial novelty, detailed at the end of this section.

Group-based Methods. These approaches are proposed to deal with cluttered scenes where detection and tracking of individuals are not reliable. Many works rely

on a foreground detector and hence do not make any distinction between individuals and groups. Following this line, a probability hypothesis density filter is used in [21] to track compact foreground areas assumed to be groups. A Kalman filter-based tracker adapted to detected groups is presented in [40]. Group splitting and merging events, not considered in the previous works, are addressed in [22] by using logical operators on foreground regions. Non-Gaussian shaped groups are estimated in [41], where foreground regions are joined by a linkage clustering algorithm. In [23], groups are modeled as nearly-regular textures using a lattice-based Markov random field.

Individual-based Methods. Some techniques of this class exploit heuristic rules [42], [43] to classify compact foreground regions as groups, individuals, or other entities. In [48], groups are detected analyzing the observed trajectories identified by the attractive or repulsive forces estimated using the Social Force Model (SFM) [49]. The SFM is also used in [29] to model the group behavior and refine the individual tracking results. Agglomerative clustering is employed in [26] to group trajectories gathered in a time-window of fixed length. The authors of [25] and [28] use pairwise spatio-temporal features extracted from the tracklets of individuals to determine a segmentation into groups. In [28], a weighted graph is created, where the weight on the edges expresses the probability of an individual being in a pairwise relation with another individual. Unfortunately, these pairwise features can only be used to predict the group and non-group state of each pair of people.

The described classes of approaches have drawbacks. The group-based techniques are limited by the assumption that groups are represented as compact regions, ignoring their structure and dynamic nature. In the individual-based methods, the performance is strongly dependent on the quality of the individual tracklets. More important, the formation of groups is seen as a mere consequence of the individuals' behavior, whereas it is widely known in sociology that groups exert important influence on how singles act [50], [51].

Joint Individual-Group-based Methods. These techniques deal with individuals and groups modeling and tracking at the same time. Many of the approaches maintain the structure of a graph in which connected components correspond to groups of individuals: in [45], stochastic differential equations are embedded in a Markov-Chain Monte Carlo (MCMC) framework implementing a probabilistic transition model for the group dynamics. However, inference with MCMC does not scale efficiently in high-dimensional state spaces. A similar framework [52] has been enriched by considering inter-group closeness and intra-group cohesion. In both cases, experiments with few targets are presented. A two-level structure for tracking using a mass-spring model is proposed in [46]: the first level deals with individual tracking, and the second level tracks individuals that are spatially coherent as a whole. Similarly, two processes are involved in [47]: the group process considers groups as atomic entities, the individual process captures how individuals move, and revises the group posterior distribution. The drawback of these methods is that they do not consider splitting and merging events, limiting their applicability.

The technique proposed in [53] models a group as a set of non-homogeneous Poisson point processes. Its advantage is that the measurement process imposes a low computational cost associated to the evaluation of the possible assignment hypotheses, but the main limitation is that the number of groups observed has to be known beforehand.

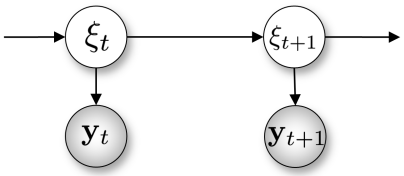
The approach in [54] barely belongs to this class of methods since groups are neither explicitly detected nor tracked. The model is built upon a conditional random field, whose potentials depend on motion, appearance and interaction between individuals. The idea is that group information is employed to better estimate the whereabouts of individuals, "sharing" state information among the members of a group. The purpose of this approach resembles that of [44], where group tracking is employed for facing individual occlusions.

In [27], a data association-based tracking is presented, where multiple individual's tracklets are clustered together in group entities, exploiting social grouping dynamics. The approach is based on a two-step optimization problem that provides both tracklet-tracklet associations (to link multiple tracklets of the same subject) and tracklet-group associations. The idea is that tracklets belonging to the same group will be related to the same individual with higher probability than the tracklets associated to different groups.

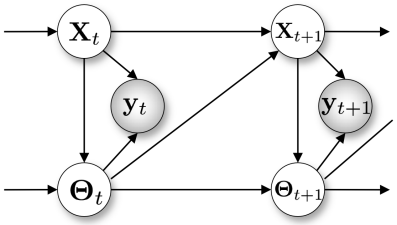
The proposed model is defined in a similar spirit to the ones presented in [29], [31], in which the tracking of individuals and their actions/collective activities are jointly modeled. Even though they share many similarities with our methods, the main differences are: 1) the methods in [29], [31] need to store the individual tracks from a large temporal window to perform inference, while we just need the position and velocity at the current time, and 2) we propose an unsupervised alternative solution to their supervised learning methods, in order to model the group configurations in the DP2-JIGT.

The proposed framework, including the DEEPER-JIGT and the DP2-JIGT, belongs to this last class of approaches and shows peculiar features. In particular:

- they can deal with an arbitrarily large (potentially infinite) and varying number of individuals and groups, unlike [45], [52], [53];
- our DPF formulation manages splitting and merging events, whereas other approaches (*e.g.*, [46], [47]) are more rigid;
- the probabilistic nature of the methods allows to either associate individuals to a single group as a hard membership, or to multiple groups as a soft assignment, unlike [25], [46].



(a) Joint state-space model.



(b) The proposed model.

Fig. 1: Models for joint individual-group tracking.

3 The Tracking Model

In this section we formally define the problem of the joint individual-group tracking, we introduce the DEEPER-JIGT based on the pre-trained classifiers and the DP2-JIGT based on online learning, and we discuss their differences.

Let us consider the general state-space model of Fig. 1a, representing the classical nonlinear discrete-time system employed for the generic object tracking. Formally, the system is defined as follows:

$$\begin{aligned} \xi_{t+1} &= f_t(\xi_t, \eta_t^{\xi}), \\ \mathbf{y}_t &= h_t(\xi_t, \eta_t^{y}) \end{aligned} \tag{1}$$

where ξ_t is the state of the system at time t, $\mathbf{y}_t$ is the observation or measurement, η_t^{ξ} and η_t^{y} are independent non-Gaussian noise variables, and f_t and h_t are nonlinear unknown functions. Eq. 1 graphically leads to the conditional link from ξ_t to ξ_{t+1} and the link from ξ_t to $\mathbf{y}_t$ in Fig. 1a.

A common inference task when using this model is to estimate at each time step t the variable ξ_t given the observations up to the current time [55], [56], that corresponds to estimate $p(\xi_t|\mathbf{y}_{0:t})$. In the case of multi-person tracking, this is analogous to estimate the position of the individuals given the image and the features that can be extracted from it. In this work, the observation $\mathbf{y}_t$ is defined using both an appearance descriptor and a person detector (see Sec. 7 for more details). In particular, detections have been simulated from the ground truth trajectories in order to decouple the evaluation of the proposed models and the choice of the detector. This choice is based on the fact that the presented framework is thought as a module of a group behavior analysis pipeline and is independent of the specific choice of person detection and target re-acquisition algorithms. We leave the study of the effect of these two modules as future work.

Let us assume that the state space can be decomposed into two subspaces that are *conditionally dependent*. The subspaces are represented by the variables $\mathbf{X}_t$ and $\mathbf{\Theta}_t$, such that $\xi_t = [\mathbf{X}_t, \mathbf{\Theta}_t]^T$. In the individual-group tracking formulation, we assume that the subspace of the individuals is $\mathbf{X}_t$ and the subspace of the groups is $\mathbf{\Theta}_t$. Where needed, the group subspace will be referred to as ${}^F\mathbf{\Theta}_t$ for the DEEPER-JIGT and ${}^N\mathbf{\Theta}_t$ for the DP2-JIGT where the superscripts F and N stand for oFfline and oNline, respectively, and discern the learning mechanism of each method. Otherwise, the symbol $\mathbf{\Theta}_t$ without superscripts generically refers to both approaches.

We rewrite the system of Eq. 1 as:

$$\begin{aligned} \mathbf{X}_{t+1} &= f_t^X(\mathbf{X}_t, \mathbf{\Theta}_t, \eta_t^X), \\ \mathbf{\Theta}_{t+1} &= f_t^{\Theta}(\mathbf{X}_{t+1}, \mathbf{\Theta}_t, \eta_t^{\Theta}), \\ \mathbf{y}_t &= h_t(\mathbf{X}_t, \mathbf{\Theta}_t, \eta_t^{y}). \end{aligned}$$

The conditional dependencies between the variables in our model are shown in Fig. 1b: the state of the individuals $\mathbf{X}_{t+1}$ depends on its previous state $\mathbf{X}_t$ and the previous state of the groups $\mathbf{\Theta}_t$. The state of the groups $\mathbf{\Theta}_{t+1}$ depends on the state of the individuals $\mathbf{X}_{t+1}$ and the previous state of the groups $\mathbf{\Theta}_t$. Finally, the observation $\mathbf{y}_{t+1}$ depends on both the state of the individuals and groups, because both of them generate the current measurements. These conditional dependencies reflect the mutual cooperation between the two subspaces.

The DEEPER-JIGT and the DP2-JIGT are two instances of the model of Fig. 1b, characterized by a different formulation of the state of the groups. In particular, the DEEPER-JIGT models explicitly the dynamics of the groups. DP2-JIGT, instead, implicitly learns their dynamics using a DPMM, thus avoiding to adopt pre-trained models.

4 Decentralized Particle Filter

In this section, we describe the general form of the DPF used to perform inference in the model of Fig. 1b. We suggest to read [32] and our supplementary material for further details.

We use the DPF to recursively estimate the posterior distribution $p(\mathbf{\Theta}_t, \mathbf{X}_{0:t}|\mathbf{y}_{0:t})$ through a *decomposition* of the joint state space in two subspaces $\mathbf{\Theta}$ and $\mathbf{X}$. In the case of two subspaces, the posterior distribution factorizes as follows:

$$p(\mathbf{\Theta}_t, \mathbf{X}_{0:t}|\mathbf{y}_{0:t}) = p(\mathbf{\Theta}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})\, p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t}) \tag{2}$$

where $\mathbf{y}_{0:t} = (\mathbf{y}_0, \ldots, \mathbf{y}_t)$ and $\mathbf{X}_{0:t} = (\mathbf{X}_0, \ldots, \mathbf{X}_t)$ represent the sequence of observations and states up to the time t, respectively. The factorization adopted by the DPF circumvents both the inefficiency and ineffectiveness of the classical particle filtering [55] when dealing with large state spaces. The main idea is to split the inference in Eq. 2 as follows:

$$p(\mathbf{\Theta}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t}) \propto p(\mathbf{\Theta}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t-1}) p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{\Theta}_t) \tag{3}$$

$$p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t}) \propto p(\mathbf{y}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t-1}) p(\mathbf{X}_t|\mathbf{X}_{0:t-1}, \mathbf{y}_{0:t-1}) \cdot p(\mathbf{X}_{0:t-1}|\mathbf{y}_{0:t-1}). \quad (4)$$

These equations highlight that the inference is online, that is, it is possible estimate $p(\mathbf{\Theta}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ and $p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t})$ given the results of inference at previous step $p(\mathbf{\Theta}_{t-1}|\mathbf{X}_{0:t-1}, \mathbf{y}_{0:t-1})$ and $p(\mathbf{X}_{0:t-1}|\mathbf{y}_{0:t-1})$, respectively. We refer to the supplementary material for all the derivations of the equations that lead to the DPF implementation.

The DPF uses sequential importance sampling [55] to perform inference for the distributions in Eq. 3 and 4 (see the supplementary material). In sequential importance sampling, a set of weighted samples (or hypotheses) are updated as soon as a new observation comes. The weights (corresponding to the importance of each sample) are first updated, given the current observation. New hypotheses are then sampled for the next time step considering their weights. The less important a hypothesis is, the less probable its survival. In DPF the sampling/weighting procedure is applied on each subspace in a nested way (see Alg. 1): first, the algorithm estimates the importance weights for $\mathbf{X}_t$ and $\mathbf{\Theta}_t$ (steps 1 and 3); second, new sample sets are generated for both the distributions using the current sample sets (steps 4 and 7). Note that there is an additional step (step 5) that enables to estimate $\mathbf{\Theta}_t$ by looking ahead to the hypotheses of $\mathbf{X}_{t+1}$, in the same spirit of [57].

More in the detail, in Step 1 of Alg. 1, the standard importance sampling formulation $(Observation \cdot Dynamics)/(Proposal)$ is applied to approximate $p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t})$. The difference with the standard framework lies in the term $\mathbf{y}_{0:t-1}$, whose presence is formally motivated by a mathematical derivation discussed in [32] and the supplementary material. Intuitively, the conditioning of $\mathbf{y}_{0:t-1}$ injects the knowledge acquired by explaining the observations $\mathbf{y}$ in the subspace $\mathbf{\Theta}$ at time $t-1$. This highlights the mutual relationship of the processes that analyze $\mathbf{X}$ and $\mathbf{\Theta}$: during the same time step, operating on $\mathbf{X}$ helps in better defining $\mathbf{\Theta}$, and across subsequent time steps, operating on $\mathbf{\Theta}$ helps $\mathbf{X}$. Step 2 is a classical re-sampling, that regularizes the distributions of the samples reducing their variance [55]. Step 3 approximates $p(\mathbf{\Theta}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ by importance sampling, assuming the group dynamics equal to the group proposal. After that, the new hypotheses for the next time instant are generated through the proposal distribution $\pi(\mathbf{X}_{t+1}|\mathbf{X}^{(i)}_{0:t}, \mathbf{y}_{0:t+1})$ (Step 4). The information encoded in that sample set is plugged into the importance sampling approximation of the posterior $p(\mathbf{\Theta}_t|\mathbf{X}_{0:t+1}, \mathbf{y}_{0:t})$ (Step 5), yielding to a second re-sampling step (Step 6) and to the final sampling of $\mathbf{\Theta}$ at time $t+1$ (Step 7).

In its classical form, the DPF produces a decomposition into low dimensional spaces of the same nature. In our case, the subspaces are still high-dimensional and particle filtering methods tend to be noisy. The effect of noise is limited here by designing the joint proposal distributions and the joint observation model (see Sec. 6) exploiting the assumptions derived by the individual-group modeling problem. In practice, all the distributions highlighted in gray in Alg. 1 have been redesigned to fit into our new context.

Algorithm 1: The DPF algorithm [32]. INPUT: samples $\{\mathbf{X}^{(i)}_t\}_{i=1,\dots,N_x}$, samples $\{\mathbf{\Theta}^{(i,j)}_t\}_{i=1,\dots,N_x, j=1,\dots,N_z}$. The superscripts (i,j) mean that for each i particle generated for describing $\mathbf{X}$ we have N_z particles for describing $\mathbf{\Theta}$. OUTPUT: importance sampling approximations of $\mathbf{X}_{t+1}$, $\mathbf{\Theta}_{t+1}$.

1. Approximate $p(\mathbf{X}_t|\mathbf{y}_{0:t})$ through the importance weights:
$$w^{(i)}_t \propto \frac{p_{N_z}(\mathbf{y}_t|\mathbf{X}^{(i)}_{0:t}, \mathbf{y}_{0:t-1}) p_{N_z}(\mathbf{X}^{(i)}_t|\mathbf{X}^{(i)}_{0:t-1}, \mathbf{y}_{0:t-1})}{\pi(\mathbf{X}^{(i)}_t|\mathbf{X}^{(i)}_{0:t-1}, \mathbf{y}_{0:t})}.$$
2. Re-sample $\{\mathbf{X}^{(i)}_t, \mathbf{\Theta}^{(i,j)}_t\}$ according to $w^{(i)}_t$.
3. Approximate $p(\mathbf{\Theta}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ through the importance weights:
$$\bar{q}^{(i,j)}_t \propto p(\mathbf{y}_t|\mathbf{X}^{(i)}_t, \mathbf{\Theta}^{(i,j)}_t).$$
4. Generate $\mathbf{X}^{(i)}_{t+1}$ according to $\pi(\mathbf{X}_{t+1}|\mathbf{X}^{(i)}_{0:t}, \mathbf{y}_{0:t+1})$.
5. Approximate $p(\mathbf{\Theta}_t|\mathbf{X}_{0:t+1}, \mathbf{y}_{0:t})$ through the importance weights
$$q^{(i,j)}_t = \bar{q}^{(i,j)}_t p(\mathbf{X}^{(i)}_{t+1}|\mathbf{X}^{(i)}_{0:t}, \mathbf{\Theta}^{(i,j)}_t).$$
6. Re-sample $\mathbf{\Theta}^{(i,j)}_t$ according to $q^{(i,j)}_t$.
7. Generate $\mathbf{\Theta}^{(i,j)}_{t+1}$ according to $\pi(\mathbf{\Theta}_{t+1}|\mathbf{X}^{(i)}_{0:t+1}, \mathbf{\Theta}^{(i,j)}_t)$.

5 Group Modeling

Recalling the definition, a self-organizing group is the result of a focused interaction, where the involved subjects spontaneously decide to be in each other's immediate presence modulating their velocity and location [17]. Based on a social signal processing approach [58], we provide a mathematical definition which embeds these characteristics in the following.

5.1 Group definition in DEEPER-JIGT and DP2-JIGT

Let $\mathbf{X}_t = \{\mathbf{x}^k_t\}^K_{k=1}$ be the joint state of the K individuals at time t and ${}^F\mathbf{\Theta}_t = \{\mathbf{z}^k_t\}^K_{k=1}$ with $\mathbf{z}^k_t \in \{0, 1, \dots, G\}$ be the joint state of the G groups[1] for the DEEPER-JIGT. We define $\mathbf{x}^k_t = [x_t, y_t, \dot{x}_t, \dot{y}_t]$ (individual positions and velocities) and $\mathbf{z}^k_t$ to be the group label for the k-th individual. As an example, let us suppose to have 5 individuals and 2 groups at time t: ${}^F\mathbf{\Theta}_t = [1, 1, 2, 2, 0]^T$ indicates that the first two individuals belong to the first group, the third and fourth individuals are in the second group, and the fifth individual is a singleton.

The DP2-JIGT exploits a group model that overcomes the hard assignment limitation of the DEEPER-JIGT by

1. Note that K and G may vary over time, but subscript t is omitted to keep the notation uncluttered.

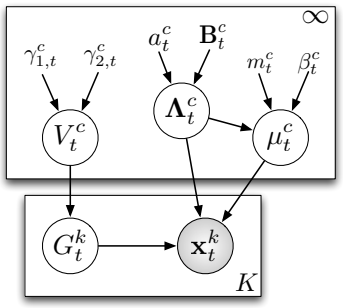


Fig. 2: Graphical model that represents a Dirichlet Process Mixture Model.

introducing a probabilistic individual-to-group assignment ${}^N\mathbf{\Theta}_t$. More formally, group modeling is addressed as a problem of mixture model fitting, where each group is defined as a mixture component in the chosen feature space. Individuals are seen as observations drawn from the mixture. The number of mixture components (*i.e.* groups) is in general unknown *a priori* and may change over time. Consequently, standard mixture models like Gaussian mixtures are not suitable since they do not deal with the dynamic inclusion and exclusion of components. Dirichlet Process Mixture Models (DPMMs) [59] have been used to overcome this limitation. DPMMs can represent mixture distributions with an unbounded number of components, where the complexity of the mixture adapts to the observed data. Moreover, we exploit a technique to perform online training of the DPMM, so that there is no need to explicitly model group events like split and merge.

The sizes of $\mathbf{X}_t$ and ${}^F\mathbf{\Theta}_t$ are fixed at time t, because new targets are deterministically initialized and added to the state space and exiting targets are deleted at each time (initialization is discussed in Sec. 7). Instead, ${}^N\mathbf{\Theta}_t$ has potentially infinite dimensions, but its size is bounded fixing the maximum number of components to C for computational reasons [60].

5.2 The Dirichlet Process Mixture Model

Each individual $\mathbf{x}^k_t$ is interpreted as an observation coming from one of the infinitely many components of the Dirichlet Process mixture. This component represents the group it belongs to.

In order to keep inference tractable, a stick-breaking representation [61] of the Dirichlet Process prior [62] is used. Such a representation is a constructive process generating an infinite set of non-negative numbers summing to 1, by sequentially sampling from a series of *Beta* distributions. The obtained sequence can be interpreted as the mixing coefficients of a set of components defining a mixture model. The graphical model associated to the generative process linking mixture components c to observations is shown in Figure 2 where, at time step t, we have K points and

$$V^c_t|\gamma^c_{1;t},\gamma^c_{2;t} \sim Beta\left(\gamma^c_{1;t},\gamma^c_{2;t}\right)$$

$$G^k_t|\{v_{1,t},v_{2,t},\ldots\} \sim Discrete\left(\rho^c_t\left(V^c_t\right)\right)$$
$$\mathbf{\Lambda}^c_t|\mathbf{B}^c_t,a^c_t \sim Wi\left(\mathbf{B}^c_t,a^c_t\right)$$
$$\mu^c_t|m^c_t,\beta^c_t,\mathbf{\Lambda}^c_t \sim \mathcal{N}\left(m^c_t,\left(\beta^c_t\mathbf{\Lambda}^c_t\right)^{-1}\right)$$
$$\mathbf{x}^k_t|G^k_t \sim \mathcal{N}\left(\mu_{G^k_t},\mathbf{\Lambda}^{-1}_{G^k_t}\right)$$

where $Discrete$ and Wi are the Discrete and Wishart distributions, respectively, $\mathbf{x}^k_t$ represents the k-th data point, G^k is an assignment variable relating each k-th data point to the mixing components, V^c and the pair $(\mu^c,\mathbf{\Lambda}^c)$ represent the parameters of the c-th mixture component in the stick-breaking construction [61], with $(\mu^c,\mathbf{\Lambda}^c)$ representing the location of the component in the parameter space and V^c defining the mixing proportions through ρ^c_t (see Sec. 6). For convenience, all the parameters of each mixture component c are grouped together as $\theta^c_t=\{\mathbf{v}^c_t,\mu^c_t,\mathbf{\Lambda}^c_t,G^c_t\}$, and ${}^N\mathbf{\Theta}_t=\left\{\theta^1_t,\theta^2_t,\ldots\right\}$.

This representation has many differences compared to the one proposed for the DEEPER-JIGT. First, the number of components, and hence groups, is unbounded. Second, the DPMM defines a probability distribution which allows to compute the probability each person has to belong to each group, rather then performing hard assignments to a single group. The introduction of a probability distribution, though, still allows to obtain the hard assignment of the DEEPER-JIGT ${}^F\mathbf{\Theta}_t$ when needed by deriving the label of the component under which the individual is most probable.

Finally, the likelihood of a data point given the model is defined as follows:

$$p(\mathbf{x}^k_t|{}^N\mathbf{\Theta}_t)=\sum_{c=1}^{\infty}\rho^c_t\cdot\mathcal{N}(\mathbf{x}^k_t|m^c_t,a^c_t\left(\mathbf{B}^c_t\right)^{-1}). \quad (5)$$

This enables us to determine the probability each person belongs to any of the groups. Such probability depends on how likely the observation associated to the person is under the Gaussian component associated to the group, and on the probability of the mixture component itself.

Note that it is always possible to generate the representation of the DEEPER-JIGT ${}^F\mathbf{\Theta}_t$ from the one of the DP2-JIGT ${}^N\mathbf{\Theta}_t$, but not the opposite. When needed, people are assigned to groups on the basis of the highest of these likelihoods:

$$\mathbf{z}^k_t=\underset{c}{\operatorname{argmax}}\left[\rho^c_t\cdot\mathcal{N}(\mathbf{x}^k_t|m^c_t,a^c_t\left(\mathbf{B}^c_t\right)^{-1})\right]. \quad (6)$$

In particular, the DP2-JIGT uses this equation to estimate the final result of the group tracker at each time step (Sec. 6.5).

5.3 Online Inference

Group dynamics is difficult to model, especially because groups tend to split and merge generating new ones. Despite that, the grouping configurations of two consecutive frames are highly correlated due to the temporal smoothness of people's trajectories. This observation is exploited in our Bayesian framework as a sequential

inference scheme, where the grouping configuration at one time step can be used as a prior belief for the next. Such sequential approach allows to account for the temporal evolution of the groups without using dynamic models (*e.g.* [33] and [63]) which are typically computationally expensive.

To pursue this, we use the sequential variational inference algorithm presented in [34] which allows us to implement online unsupervised learning of group structures. The algorithm, based on [60], relies on a truncation of the stick-breaking construction. By performing such truncation appropriately (*i.e.* keeping a large number of components), the introduced approximation is negligible [64] and only the components supported by data are actually used by the model. Single updates are performed for each frame and the obtained approximate posterior over the mixture model is used as a prior for the grouping configuration in the following frame. This is achieved by sequentially updating the parameters of the model (γ_1, γ_2, a, B, m, β) (see Figure 2) estimated at time $t-1$ (prior for time t) using the data observed at time t.

We constrain the possible mixture components (*i.e.* group configurations) inferred by the learning algorithm, considering elements of proxemics [65], which assume that people tend to unconsciously organize the space around them in concentric zones with different degrees of intimacy. The shorter the distance between two persons, the higher the degree of intimacy. Thus, we define a limit distance ($r = 2$ meters [10], beyond which two individuals can be considered not to be interacting with high probability. When this limit is not respected, the corresponding mixture component is discarded and a new one is initialized in a region of the space which is badly modeled by the mixture distribution.

6 Joint Individual-Group Tracking

This section describes how the probability distributions highlighted in gray in Alg. 1 are designed for the joint individual-group tracking problem. In particle filtering, distributions may have an analytic form, or they can be approximated by particles' sets. The latter is preferred when dealing with arbitrarily complex distributions. Analytic functions may enable a closed-form solution of the inference problem, however, their simplicity (*e.g.*, Gaussian distribution of the Kalman filter) reduces the expressiveness of the tracking posterior. Therefore, practical implementations often employ both in order to have a posterior distribution that is able to represent rather complex functions.

As a notation reminder, symbol $\mathbf{\Theta}_t$ addresses generally the group subspace, while ${}^F\mathbf{\Theta}_t$ indicates its instantiation for the DEEPER-JIGT, and ${}^N\mathbf{\Theta}_t$ corresponds to the group variable proper of the DP2-JIGT.

6.1 Individual Proposal $\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{0:t+1})$

This distribution models the dynamics of the individuals. Inspired by [66], we adopt the notion of composite proposal, incorporating two sources of information:

$$\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{0:t+1}) = \pi(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{y}_{t+1}) = \\ = \alpha\, \pi_{\text{dyn}}(\mathbf{X}_{t+1}|\mathbf{X}_t) + (1-\alpha)\, \pi_{\text{det}}(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{y}_{t+1}),$$

where we assume Markovianity between the $\mathbf{X}$s and conditional independence with respect to the observations $\mathbf{y}_{0:t}$. We adopt a locally-linear dynamics with Gaussian noise:

$$\mathbf{x}_{t+1}^k = A\mathbf{x}_t^k + \eta^x \quad \text{with } A = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where T is the sampling interval and $\eta^x \sim \mathcal{N}(\mathbf{0}, \Sigma^k)$. Therefore, $\mathbf{x}_{t+1}^k \sim \mathcal{N}(A\mathbf{x}_t^k, \Sigma^k)$, that is easy to evaluate and sample from. We have:

$$\pi_{\text{dyn}}(\mathbf{X}_{t+1}|\mathbf{X}_t) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_{t+1}^k | A\mathbf{x}_t^k, \Sigma^k) \tag{7}$$

that is, a multivariate Gaussian distribution with block-diagonal covariance matrix: $\text{diag}(\Sigma^1, \Sigma^2, \ldots, \Sigma^K)$. We assumed that $\Sigma^k = \Sigma$ for each $k = 1, \ldots, K$.

The second term $\pi_{\text{det}}(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{y}_{t+1})$ generates hypotheses in those region of the state space where it is more probable to find a person. In practice, we are using an informative proposal that searches for detected people. When a person is detected the tracker will be more reliable[2]. The distribution is defined as a multivariate Gaussian distribution centered in the detections associated to each target with covariance matrix Σ (same as Eq. 7). The parameter $\alpha = 0.5$ is fixed for all the experiments. This distribution is the same for the DEEPER-JIGT and the DP2-JIGT.

6.2 Joint Observation Distribution $p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{\Theta}_t)$

The joint observation distribution is defined to account for both the appearance and the group membership contributions as follows:

$$p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{\Theta}_t) \propto g_{\text{app}}(\mathbf{y}_t, \mathbf{X}_t, \mathbf{\Theta}_t) \cdot g_{\text{mem}}(\mathbf{y}_t, \mathbf{X}_t, \mathbf{\Theta}_t).$$

In both the DEEPER-JIGT and the DP2-JIGT, we assume that only the individuals contribute in the appearance component. Therefore, we define:

$$g_{\text{app}}(\mathbf{y}_t, \mathbf{X}_t, \mathbf{\Theta}_t) = \prod_{k=1}^{K} e^{-\lambda_y\, d_y(q(\mathbf{y}_t, \mathbf{x}_t^k), \tau^k)},$$

where $q(\mathbf{y}_t, \mathbf{x}_t^k)$ extracts a descriptor from the current bounding box in the image given by $\mathbf{x}_t^k$, τ^k is the descriptor of the template of the k-th individual and d_y is a distance between descriptors. It is easy to notice that the appearance component is defined as the standard

2. The detections have been simulated with false positive and negative rates of 20%, because performing detection in such scenarios is challenging and it is out of the purposes of this work. Data association between tracks and detections is based on nearest neighbors, but more advanced techniques [67] can easily be implemented.

template-based technique [68], widely used in the particle filtering approaches [69]. In this work, we used the Bhattacharyya distance between RGB color histograms and the template is never updated. Please note that the generality of our formulation enables us to use more sophisticated techniques and appearance models of the target.

The membership component is defined in different ways for the DEEPER-JIGT and the DP2-JIGT, because it depends on how groups are modeled. In the DEEPER-JIGT, we assumed that $\mathbf{y}_t$ do not give any contribution, therefore the definition is the following:

$$g_{\text{mem}}(\mathbf{y}_t, \mathbf{X}_t, {}^F\mathbf{\Theta}_t) = e^{-\lambda_{cl}\, d_{cl}({}^F\mathbf{\Theta}_t, \mathbf{X}_t)} \prod_{g=1}^{G} \mathcal{N}(S_t^g | \mu, \sigma),$$

where $d_{cl}({}^F\mathbf{\Theta}_t, \mathbf{X}_t)$ is a cluster validity measurement, such as the Davies-Bouldin index [70] and S_t^g is the size of the g-th group in ${}^F\mathbf{\Theta}_t$. The second term of the membership component penalizes the hypotheses of having too-large groups (in the experiments, $\mu = 1$ and $\sigma = 1.5$). However, such measurements are usually heuristics and they do not ensure that a model is better than another in absolute terms.

Instead, we define the membership function for the DP2-JIGT in a more elegant way by using directly Eq. 5 as follows:

$$g_{\text{mem}}(\mathbf{y}_t, \mathbf{X}_t, {}^N\mathbf{\Theta}_t) = \prod_{k=1}^{K} \sum_{c=1}^{C} \rho_t^c \cdot \mathcal{N}(\mathbf{y}_t^{x^k} | m_t^c, a_t^c\, (\mathbf{B}_t^c)^{-1}),$$

where C is the truncation level of the mixture [60],

$$\rho^c = \begin{cases} v^c & \text{if } c = 1 \\ v^c \cdot \prod_{j=1}^{c-1}(1 - v^j) & \text{if } c > 1 \end{cases}$$

being

$$v^c = \begin{cases} \frac{\gamma_1^c}{\gamma_1^c + \gamma_2^c} & \text{if } c < C \\ 1 & \text{if } c = C \end{cases}.$$

and $\mathbf{y}_t^{x^k}$ is the detection associated to the kth individual. The membership function encourages the hypotheses that fit well with the mixture components and therefore it is more likely to obtain compact groups.

6.3 Joint Individual Distribution $p(\mathbf{X}_{t+1} | \mathbf{X}_{0:t}, \mathbf{\Theta}_t)$

This distribution models the dynamics of the individual, taking into account the presence of the group. The idea is to map ${}^N\mathbf{\Theta}_t$ to ${}^F\mathbf{\Theta}_t$ through Eq. 6, using the same distribution for the DP2-JIGT and the DEEPER-JIGT.

Let us define the dynamics as follows:

$$\mathbf{x}_{t+1}^k = \mathbf{x}_t^k + B\mathbf{g}_t^k + \eta^x \tag{8}$$

where

$$B = \begin{bmatrix} 0 & 0 & T & 0 \\ 0 & 0 & 0 & T \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \; \mathbf{g}_t^k = \frac{\sum_{l=1}^{K} \mathbf{x}_t^l\, \mathbb{I}(\mathbf{z}_t^k == \mathbf{z}_t^l)}{\sum_{l=1}^{K} \mathbb{I}(\mathbf{z}_t^k == \mathbf{z}_t^l)}$$

$\mathbb{I}(\cdot)$ is the indicator function and $\mathbf{g}_t^k$ is the position and velocity of the group the k-th individual belongs to. Note that the matrix B selects only the velocity vector of $\mathbf{g}_t^k$, discarding the positional information. This encourages individuals in the same group to have similar dynamics. Similarly to Eq. 7, the resulting probability distribution is:

$$p(\mathbf{X}_{t+1} | \mathbf{X}_{0:t}, \mathbf{\Theta}_t) = p(\mathbf{X}_{t+1} | \mathbf{X}_t, \mathbf{\Theta}_t) = \\ = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_{t+1}^k | \mathbf{x}_t^k + B\mathbf{g}_t^k, \Sigma).$$

6.4 Joint Group Proposal $\pi(\mathbf{\Theta}_{t+1} | \mathbf{X}_{0:t+1}, \mathbf{\Theta}_t)$

In this case, the two approaches lead to completely different definition of this distribution but share the Markov assumption: $\pi(\mathbf{\Theta}_{t+1} | \mathbf{X}_{0:t+1}, \mathbf{\Theta}_t) = \pi(\mathbf{\Theta}_{t+1} | \mathbf{X}_{t+1}, \mathbf{\Theta}_t)$

The idea followed by the DEEPER-JIGT is to use a surrogate distribution over the possible events that may happen to a group (namely merge, split, and none). The surrogate distribution is thus easier to sample from than the original proposal. The joint group proposal for the DEEPER-JIGT is defined as:

$$\pi({}^F\mathbf{\Theta}_{t+1} | \mathbf{X}_{t+1}, {}^F\mathbf{\Theta}_t) = f(\prod_{g=1}^{G} \pi(e_{t+1}^g | \mathbf{X}_{t+1}, g_t, g_t'), {}^F\mathbf{\Theta}_t) \tag{9}$$

where $\pi(e_{t+1}^g | \mathbf{X}_{t+1}, g_t, g_t')$ is the surrogate distribution defined on the set of events $e^g \in \{\text{Merge}, \text{Split}, \text{None}\}$ related to the g-th group and g' is the group associated to the g-th group using the distance between their centroids.

The surrogate distribution is learned offline, via multinomial logistic regression. To this end, a set of possible scenarios containing events have been simulated and labeled [33]. We use as features 1) the inter-group distance between g and the nearest group g', considering their positions and sizes (d_{KL}, symmetrized Kullback-Leibler distance between Gaussians) and velocities (d_v, Euclidean distance), and 2) the intra-group variance between the positions of the individuals in the g-th group (d_{intra}). Thus, the input of the multinomial logistic regression is a 6-dimensional vector, *i.e.* $(d_{KL}, d_v, d_{\text{intra}})$ for time t and $t+1$. More details are provided in the supplementary material.

The DP2-JIGT learns the proposal distribution in an online fashion. In detail, let us assume that the DPMM ${}^N\tilde{\mathbf{\Theta}}_t$ is trained using the individual state estimate (not the particles) up to the current frame, so that only one model is kept at each time step. To sample the new hypotheses set at time $t+1$, a copy of the model ${}^N\mathbf{\Theta}_t^{(i)}$ is instantiated for each new particle as the DPMM learned to date. The model ${}^N\tilde{\mathbf{\Theta}}_t^{(i)}$ is updated using the i-th hypothesis $\mathbf{X}_{t+1}^{(i)}$ in accordance with the method discussed in Sec. 5.3 thus generating N different models ${}^N\mathbf{\Theta}_{t+1}^{(i)}$, $i = \{i, \ldots, N\}$.

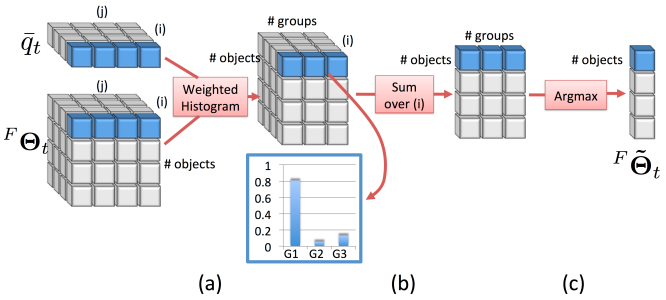


Fig. 3: Computation of the state estimate $^F\tilde{\Theta}_t$ that deals with discrete labels.

Then, M samples are generated from each model, that is, $^N\Theta_{t+1}^{(i,j)}$, $i = \{i, \dots, N\}, j = \{1, \dots, M\}$. In order to perform this second sampling step, new grouping hypotheses are formulated on the basis of what has been learned by the model. To this end, a new mixture hypothesis is assigned to each particle by performing ancestral sampling from the graphical model in Fig. 2, for each of the mixture components. In practice, the parameters of component c are obtained by fixing the hyperparameters γ_1^c, γ_2^c, $\mathbf{B}^c$, a^c, m^c, β^c to their current estimate and sampling in a top-down manner the random variables in the graphical model, sampling children once all their parents have been sampled.

The characteristics of the joint group proposal of the DEEPER-JIGT are that it relies on offline training of a classifier that detects the events and is based on a supervised learning technique, *i.e.* the group events should be annotated beforehand, often on simulated/synthetic training data. On the other hand, the DP2-JIGT performs online learning without explicitly modeling the group events and is unsupervised, *i.e.* labeled data are not required.

6.5 State Estimate

The last step is related to the estimation of the most likely joint state given the observations. The joint state is usually defined as the expected value of the state under a certain distribution [55], that is, $\mathbb{E}_{p(\mathbf{X}_t|\mathbf{y}_{0:t})}[\mathbf{X}_t]$ and $\mathbb{E}_{p(\mathbf{\Theta}_t|\mathbf{X}_t,\mathbf{y}_{0:t})}[\mathbf{\Theta}_t]$.

The expected value of the individual state $\mathbf{X}_t$ can be estimated as $\tilde{\mathbf{X}}_t = \sum_{i=1}^{N_x} w_t^{(i)} \mathbf{X}_t^{(i)}$, given the empirical approximation of $p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t})$. Instead, the expected value of the individual state $\mathbf{\Theta}_t$ cannot be computed directly, in case of the DP2-JIGT, because it implies to average group model hypotheses and thus the associations between the components of different hypotheses. To overcome this problem, we first map each model $^N\mathbf{\Theta}_t$ to the hard assignment of individuals $^F\mathbf{\Theta}_t$ using Eq. 6. Then, the same procedure to estimate the group state can be used for the DEEPER-JIGT and the DP2-JIGT.

We compute a distribution over the possible labels as depicted in Fig. 3. Starting from the matrices $^F\mathbf{\Theta}_t$ and $\bar{q}_t$, we compute the following distribution for the k-th individual as a weighted histogram:

$$\mathrm{Wh}^{k,(i,g)} = \sum_{j=1}^{N_z} \bar{q}_t^{(i,j)} \mathbb{I}(\mathbf{z}_t^{k,(i,j)} == g).$$

This gives a similar representation of the sum over j but it considers labels g (step (a) in Fig. 3). Then, each $\mathrm{Wh}^{k,(i,g)}$ is summed over i (step (b) in Fig. 3), and we take the maximum likelihood estimate of the association between groups and individuals to obtain $^F\tilde{\mathbf{\Theta}}_t$ (step (c) in Fig. 3).

7 Experiments

The proposed methods were tested and evaluated on two challenging datasets: Friends Meet[3] (FM) [33] and BIWI[4] [71]. In order to assess tracking performance and investigate the effects of the mutual support of the group and individual tracking processes, we adapted the standard evaluation metrics proposed for individual tracking. To this end, bounding boxes have been replaced by the convex hull enveloping all the individuals inside the group in all the metrics.

In our experiments, we decided to decouple the evaluation of our models from the specific choice of person detector in order to assess the theoretical value of the conditional dependency relations introduced between groups and individuals. To this end, we simulated the person detector by generating detections from the ground truth with a false positive and negative rates of 20% and adding spatial Gaussian noise. We leave the study of the effect of state-of-the-art detectors as future work.

The proposed DEEPER-JIGT and DP2-JIGT differ from other state-of-the-art approaches, since are able to manage a set of conditions (variable number of individuals and groups, split/merge events) never managed jointly by other studies. This, together with the lack of evaluation metrics proper for group tracking, make the comparative tests hard. To deal with this issue, we compare our two approaches with different versions that 1) exhibit a simplified dynamics for both groups and individuals, relaxing the mutual influences, 2) treat the problem using two separate trackers, one for the groups and one for the individuals, and 3) mimic generic multi-person trackers, which deal with groups as they were atomic entities, resembling approaches like [1], [2], [3], [4], [5].

We prove that: 1) both the DEEPER-JIGT and the DP2-JIGT perform joint tracking with some differences between them, 2) methods that resemble multi-person trackers perform poorly on groups, 3) our idea allows to obtain an optimal compromise between individual and group tracking performance.

3. Downloadable at http://goo.gl/cFXCG
4. Downloadable at http://www.vision.ee.ethz.ch/datasets/index.en.html

7.1 Datasets and Evaluation Metrics

The FM dataset [33] is composed by 53 sequences (for a total of 16286 frames) partitioned in a *synthetic* and a *real* dataset. In the following, the generic term *event* is used to refer to group splits, group merges and the entrance or exit of a group in the scene. The synthetic set presents objects with a simple and stable appearance which are moving according to a non-linear dynamics (see some examples in the supplementary material). It contains 18 easy sequences with 1-2 group events and 4-10 individuals, and 10 more challenging sequences with 10-16 individuals involved in multiple events.

The *real* set focuses on an outdoor area where individuals usually meet during coffee breaks (see examples in the supplementary material). This area has been recorded and annotated by an expert for one month. The expert reported the events that appeared more frequently, building a screenplay where these events are summarized in order to limit the dataset size. Then, the screenplay was played by students and employees, resulting in 15 sequences of different length (between 30 sec. to 1.5 minutes) which were considered to be sufficiently realistic by the expert. The sequences contain from 3 to 11 individuals, they are all annotated with individual and group information.

The subset of the sequences [33] that does not contain queues of individuals was selected. Queues were excluded from the experiments because we model the joint individual-group tracking of self-organizing groups, as we discussed in Sec. 1. In contrast with the definition we employed, queues are defined as the typical example of a *circumstantial* group, where unplanned and often temporary group formations arise, due to external forces that bring people together [16].

To further assess the performance of the proposed framework, we considered also the BIWI dataset [71]. This represents an outdoor scenario where people walk following the same direction from a source to a destination in a limited interval of time. Split and merge events in this case are rare, therefore the dataset is not the ideal benchmark for our method, since some of its capabilities cannot emerge properly.

The proposed methods are evaluated in terms of individual and group accuracy using standard tracking metrics, such as False Positive (FP) and False Negative (FN) rates [72] for detection, Mean Square Error (MSE) of the estimated positions and its standard deviation, Multi-Object Tracking Precision (MOTP) and Accuracy (MOTA) [73] for tracking and id-switch (ID) [74]. In all these metrics, intersection operations among bounding boxes of individuals translate naturally in intersections among convex hulls of groups. We also introduced the Group Detection Success Rate (GDSR) as the rate of the correctly detected groups. In this case we consider that a group is *correctly detected* if at least the 60% of its members are detected.

TABLE 1: Results on the *synthetic* FM dataset excluding the queue sequences (see text for the details). Individual tracking (column 2), group detection (columns 3-5) and group tracking (column 6-7). For MSE and MOTP (in pixels), the lower the better. In bold, the best results only comparing the tracking algorithms.

	MSE [px] (std)	1-FP	1-FN	GDSR	MOTP [px]	MOTA
DP2-JIGT	**1.75** (**4.76**)	**93.98%**	**91.28%**	**86.91%**	**16.72**	**71.57%**
DEEPER-JIGT	2.28 (5.42)	93.12%	81.01%	78.18%	18.16	53.43%
VAR1	2.29 (5.98)	93.30%	79.88%	76.76%	19.34	52.19%
VAR2	4.00 (13.09)	81.20%	47.06%	45.22%	168.19	28.61%
VAR3	2.66 (9.21)	65.24%	20.05%	15.75%	442.81	4.00%
DPMM det. [34]	–	94.30%	91.89%	88.57%	–	–

7.2 Results: Synthetic Scenarios

The first evaluation on the synthetic part of the FM dataset is focused on comparing the DP2-JIGT, the DEEPER-JIGT and variants that relax the conditional probabilities of the DEEPER-JIGT. We show that the conditional dependencies between the two subspaces are required to perform individual-group tracking.

The variants of the DEEPER-JIGT are called: VAR1, VAR2 and VAR3. VAR1 assumes $p(\mathbf{X}_{t+1}|\mathbf{X}_t, {}^F\mathbf{\Theta}_t) = p(\mathbf{X}_{t+1}|\mathbf{X}_t)$, inhibiting the contribution of the group on the dynamics of the individual. This is done by dropping out the term $B\mathbf{g}_t^k$ in Eq. 8. VAR2 is the same as VAR1, assuming in addition $\pi({}^F\mathbf{\Theta}_{t+1}|\mathbf{X}_{t+1}, {}^F\mathbf{\Theta}_t) = \pi({}^F\mathbf{\Theta}_{t+1}|{}^F\mathbf{\Theta}_t)$ (Eq. 9), that is, suppressing the influence of individual states on the group configurations evolution. In practice, instead of sampling from the surrogate distribution of events, we sampled from the combinatorial space of possible configurations of the group hypothesis, supposing they are uniformly distributed. Going from the DEEPER-JIGT to VAR2, we can notice that $\mathbf{X}$ and $\mathbf{\Theta}$ become independent from each other, and thus sampling is performed independently in each state space. The only joint contribution remaining is in the observation model. Finally, VAR3 is VAR2 with the assumption $p(\mathbf{y}_t|\mathbf{X}_t, {}^F\mathbf{\Theta}_t) \propto g_{\text{app}}(\mathbf{y}_t, \mathbf{X}_t, {}^F\mathbf{\Theta}_t)$, removing the contribution of groups, *i.e.*, $g_{\text{mem}}(\mathbf{y}_t, \mathbf{X}_t, {}^F\mathbf{\Theta}_t) = 1$. This way, the model considers how individuals are tracked, but not how well they fit the current group configuration hypotheses. In practice, this variant separates individual tracking from group tracking in two different particle filters. Moreover, we reported the results using the DPMM detector [34] applied to the individual tracks given by the DEEPER-JIGT (last row in Table 1).

The statistics reported in Table 1 show that the DP2-JIGT and the DEEPER-JIGT outperform VAR3, VAR2, VAR1. This suggests that each distribution introduced in the model contributes to obtain better accuracy in group detection and tracking.

Moreover, considering the MSE and its standard de-

TABLE 2: Results on the real FM dataset excluding the queue sequences (see the text for the details). Group detection (columns 2-4) and group tracking (column 5-6). For MOTP (in meters), the lower the better.

	1-FP	1-FN	GDSR	MOTP [m]	MOTA
DP2-JIGT	**97.81%**	**97.54%**	**94.65%**	0.92	**73.85%**
DEEPER-JIGT	95.72%	89.99%	85.78%	**0.87**	65.18%
VAR3	75.02%	32.51%	21.53%	3.12	3.14%
Data ass. (baseline)	97.92%	98.21%	94.89%	1.02	70.35%

viation (first column, in brackets) we can notice that the tracker of the individuals exploits the information provided by the tracker of groups. This is even clearer when considering the DP2-JIGT, where the best MSE is obtained. It is also easy to note that the DP2-JIGT outperforms the DEEPER-JIGT over all the performance statistics. This finding strongly promotes the use of the online learning method embedded into the DPF for individual-group tracking in cocktail-party scenarios.

The proposed approaches show lower detection performance with respect to the DPMM detector (last row). This happens because they try to maintain the individual and group labels consistently during the sequence, employing a dynamic model that in some cases fails to predict adequately the whereabouts of the individuals. The DPMM detector, despite its better performance, is not able to deal with the temporal evolution of grouping formations.

7.3 Results: Real Scenarios

Given the difficulties of real scenarios, the tracker of the individuals can lose some targets because of occlusions, low resolution, and due to the fact that their appearance model is not updated over time. Particular attention has to be paid to the initialization issue. When an individual target is lost (distance above 0.6 meters) we re-initialize the individual track, using the detections simulated by the ground truth (as described in Sec. 7). This allows us not to use any specific algorithm to perform re-initialization and re-acquisition of the targets. The average re-initialization rate per track is reported in Table 3. It is worth noticing that there exist many re-initialization strategies that can be adopted in this step, see [1] as an example.

The statistics on the real FM dataset, reported in Table 2, are consistent with the previous experimental findings. The DEEPER-JIGT performs better than VAR3, and the DP2-JIGT outperforms the DEEPER-JIGT, especially in terms of false negative rate and the GDSR. This suggests that there are less false negatives and that we have a more accurate localization of the groups when detected. In terms of tracking accuracy (last column), the DP2-JIGT is still better than the DEEPER-JIGT but it less precise (slightly higher MOTP).

To further assess the performance of the proposed models, we compared them with a group tracking method based on data association that models groups

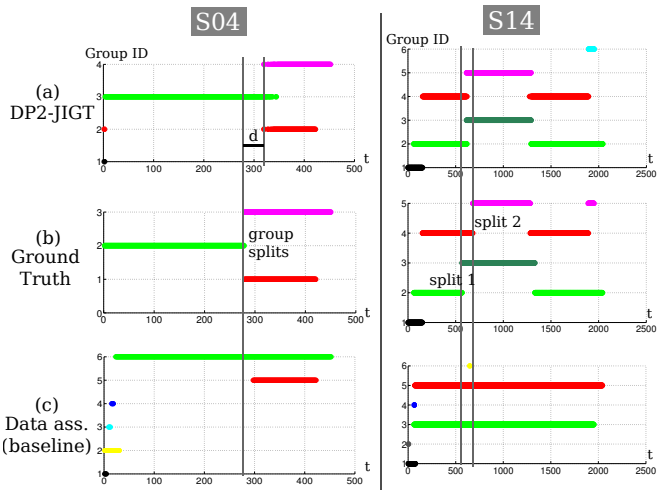


Fig. 4: Qualitative results on the real FM dataset (S04 and S14) comparing the DP2-JIGT and the data association-based baseline. Better printed in color.

TABLE 3: Results of individual tracking for the real FM dataset excluding the queue sequences (see the text for the details).

	1-FP	1-FN	MSE [px]	MOTP [px]	Re-init	ID
DP2-JIGT	81.25%	78.11%	0.25	0.71	3.2%	156
DEEPER-JIGT	82.87%	79.82%	0.24	**0.71**	3.3%	148
VAR3	**88.12%**	**84.05%**	**0.22**	0.72	**1.6%**	**132**

as atomic entities, acting as a generic multiple-target tracker [1], [2], [3], [4], [5]. This data association-based tracker uses the detections of groups given by the DPMM detector [34]. Group tracking is performed by associating the groups at time t with the groups at time $t-1$ through nearest neighbor on the position-velocity vector. A detailed comparison on the different scenarios is reported in the supplementary material.

The statistics reported on Table 2 (last row) show that while the detection accuracy measures obtained with the described data association technique (FP, FN, and GDSR) are comparable with those obtained by DP2-JIGT, the DP2-JIGT outperforms it in terms of tracking (MOTP and MOTA). This is due to the fact that the data association-based tracker is not able to deal correctly with splitting and merging events. This is especially highlighted by the results of two sequences reported in Fig. 4 (S04 and S14). In the rows, we reported the results of the DP2-JIGT, the ground truth and the data association baseline. Each graph has the group ids on the y axis and the time on the x axis. The two graphs show that the baseline is not able to assign a new group id whenever a split occurs (e.g., time 275 S04, time 510 S14). On the other hand, the DP2-JIGT is able to correctly detect the events and the new groups with a small delay.

We also analyzed the individual tracking performance in the FM dataset, reported in Table 3. One can notice that the DP2-JIGT, DEEPER-JIGT and VAR3 trackers are comparable in terms of MSE and MOTP, while VAR3

TABLE 4: Group results on the BIWI dataset.

	1-FP	1-FN	GDSR	MOTP [m]	MOTA
DP2-JIGT	37.66%	**89.43%**	51.86%	0.47	22.94%
DEEPER-JIGT	53.77%	78.00%	**53.59%**	**0.44**	**29.43%**
VAR3	**60.55%**	51.57%	29.60%	1.03	9.58%

outperforms the DP2-JIGT and DEEPER-JIGT in terms of the other measures (FP, FN, re-init, and ID). This is due to the fact that VAR3 models the individual tracking and the group tracking as two separated processes. On one hand, the individual component is not influenced by the group tracking, thus favoring a more accurate estimate of individuals. On the other hand, this independence causes poor group tracking performance, because group tracking does not consider the individual component. In practice, we observed that VAR3 produces either big group estimates (low FP rate) or no group estimate at all (low FN rate). Since the group estimates of VAR3 vary consistently over time, new group identifiers are generated, resulting in high group MOTP and low MOTA (see Table 2). Therefore, when considering the results of of Table 2 and Table 3 jointly, the proposed DP2-JIGT and DEEPER-JIGT provide the best compromise between individual and group tracking performance. In other words, a small loss in the individual tracking performance corresponds to a consistent advantage in group tracking.

The BIWI dataset [71] presents different challenges for group tracking: first of all, group dynamics is poor, *i.e.* group events are very rare, and secondly group events are not annotated in the ground truth. Unfortunately, the literature lacks of other datasets where group events are present and annotated for the tracking purpose (*i.e.* keeping consistence of group labels across different frames). The two sequences of the BIWI have individuals who walk alone in one direction and stay in the field of view of the camera for few frames.

We carried out a set of experiments showing the results of the DP2-JIGT, the DEEPER-JIGT and one of its variants (VAR3) using this dataset. The results were computed for the annotated frames of the sequences where the ground-truth is available from [71].

The results are reported in Table 4. The DP2-JIGT outperforms the DEEPER-JIGT in terms of false negative rate, which means that is less conservative than the DEEPER-JIGT which generates less groups. Despite this benefit, the DP2-JIGT pays in terms of false positive and tracking accuracy. The main reason for this shortcoming is that the life span of a group in this dataset is really limited to few tens of frames. The online DPMM needs a bootstrap period to learn plausible configurations of new groups, that, in this case, is greater than their life span. This generates more false detections and tracks. One possible solution to this problem could be to impose a constraint on the size of the group embedded into the DPMM algorithm. We leave this point as a future extension of the work.

Some qualitative results that compares the DEEPER-JIGT and the DP2-JIGT are reported in Fig. 3 of the additional material and the video at http://youtu.be/TOYm060sZDc. In particular, an advantage of the DP2-JIGT is that it is able to initialize groups faster than the DEEPER-JIGT. This is due to the fact that the DEEPER-JIGT tries to merge pairs of individuals and/or groups, while the DP2-JIGT uses all the data simultaneously.

We also noticed that the DEEPER-JIGT tends to merge groups with singleton and sometimes with other groups even if they are far away (*e.g.*, S07 $t = 202$ and $t = 366$), while the DP2-JIGT uses the social constraint to avoid it.

The advantage of the DP2-JIGT over the DEEPER-JIGT is due to the fact that the sequences are long enough to learn online the way groups evolve, bringing the DP2-JIGT to generate better hypotheses during tracking. On the other hand, the DEEPER-JIGT is preferred when the individuals stay on the scene for a limited period of time and when the scenario is particularly crowded (*e.g.*, the BIWI sequence "eth", see last two columns, last two rows in Fig. 3 of the additional material). These findings suggest that depending on the monitored scenario one can prefer to use either the DEEPER-JIGT or the DP2-JIGT. In particular, in dynamical videos where individuals and groups stay for a very limited time span, the DEEPER-JIGT is preferred since it does not need any bootstrap learning phase. Instead, in other scenarios where individuals often engage in interplay for a longer period of time, the DP2-JIGT appears to be the best choice.

8 Conclusions

This study promotes the innovative idea of simultaneous modeling of groups and individuals in videos. This allows information sharing between the two tracking instances in a mathematical sound framework through a Decentralized Particle Filter. We devised two variants of this core idea that allow to have an optimal compromise between individual and group tracking performance. A first version which is built upon pre-trained classifiers that model how groups evolve in the scene and a second version capable of learning group dynamics in an online fashion, thanks to a Dirichlet Process Mixture Model.

The proposed framework has several aspects that could be extended. First, the framework should be tested using a state-of-the-art person detector as [7]. Then, a second-order dynamical model and an online-learned appearance descriptor could improve the individual tracking performance. Our approach is general enough to allow the embedding of these methods in the presented framework separately and in a modular fashion. Regarding the groups, diminishing the bootstrap time of the DP2-JIGT may lead to a more effective tracker, definitely outperforming the offline learning of the DEEPER-JIGT.

REFERENCES

[1] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1820–1833, 2011.

[2] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2012.

[3] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, june 2012, pp. 1918 –1925.

[4] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multi-target tracking." *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[5] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1409–1422, 2012.

[6] J. Pilet, C. Strecha, and P. Fua, "Making background subtraction robust to sudden illumination changes," in *European Conference on Computer Vision (ECCV)*. Springer, 2008, pp. 567–580.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2010.

[8] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Comput. Vis. Image Underst.*, vol. 117, no. 2, pp. 130–144, Feb. 2013.

[9] Q. Le, W. Zou, S. Yeung, and A. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, june 2011, pp. 3361 –3368.

[10] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1743–1759, 2009.

[11] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2064–2070, Oct. 2012.

[12] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 1235–1242.

[13] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *European Conference on Computer Vision (ECCV)*, 2008, pp. 1–14.

[14] D. R. Forsyth, *Group dynamics*. Cengage Learning, 2010.

[15] E. Goffman, "Encounters: Two studies in the sociology of interaction." 1961.

[16] H. Arrow, J. McGrath, and J. Berdahl, *Small Groups as Complex Systems: Formation, Coordination, Development, and Adaptation*. SAGE Publications, 2000.

[17] E. Goffman, *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Free Press, 1966.

[18] W.-S. Zheng, S. Gong, and T. Xiang, "Group association: Assisting re-identification by visual context," in *Person Re-Identification*. Springer, 2014, pp. 183–201.

[19] M. A. Yousuf, Y. Kobayashi, Y. Kuno, A. Yamazaki, and K. Yamazaki, "Development of a mobile museum guide robot that can configure spatial formation with visitors," in *Intelligent Computing Technology*. Springer, 2012, pp. 423–432.

[20] A. Kendon, *Conducting Interaction: Patterns of behavior in focused encounters*, C. U. Press, Ed., 1990.

[21] Y.-D. Wang, J.-K. Wu, A. A. Kassim, and W.-M. Huang, "Tracking a variable number of human groups in video using probability hypothesis density," in *International Conference on Pattern Recognition (ICPR)*, 2006.

[22] G. Gennari and G. Hager, "Probabilistic data association methods in visual tracking of groups," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[23] W.-C. Lin and Y. Liu, "A lattice-based MRF model for dynamic near-regular texture tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 777–792, 2007.

[24] S. Pellegrini, A. Ess, and L. V. Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 452–465.

[25] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg, "Who are you with and where are you going?" in *IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, 2011.

[26] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1003–1016, 2012.

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

[28] M. Chang, N. Krahnstoever, and W. Ge, "Probabilistic group-level motion analysis and scenario recognition," in *IEEE ICCV*, 2011.

[29] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: modeling social and grouping behavior on a linear programming multiple people tracker," *1st Workshop on Modeling, Simulation and Visual Analysis of Large Crowds*, 2011.

[30] S. Khamis, V. I. Morariu, and L. S. Davis, "A flow model for joint action recognition and identity maintenance," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2012, pp. 1218–1225.

[31] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 215–230.

[32] T. Chen, T. Schon, H. Ohlsson, and L. Ljung, "Decentralized particle filter with arbitrary state decomposition," *IEEE Trans. on Signal Processing*, vol. 59, no. 2, pp. 465–478, 2011.

[33] L. Bazzani, V. Murino, and M. Cristani, "Decentralized particle filter for joint individual-group tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[34] M. Zanotto, L. Bazzani, M. Cristani, and V. Murino, "Online bayesian non-parametrics for social group detection," in *British Machine Vision Conference (BMVC)*, 2012.

[35] J. C. S. Jacques, A. Braun, J. Soldera, S. R. Musse, and C. R. Jung, "Understanding people motion in video sequences using voronoi diagrams: Detecting and classifying groups," *Pattern Analysis Applications*, vol. 10, no. 4, pp. 321–332, 2007.

[36] L. Bazzani, M. Cristani, G. Pagetti, D. Tosato, G. Menegaz, and V. Murino, "Analyzing groups: a social signaling perspective," in *Video Analytics for Business Intelligence*, ser. Studies in Computational Intelligence. Springer-Verlag, 2012.

[37] A. Fathi, "Social interactions: A first-person perspective," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1226–1233.

[38] M. Cristani, L. Bazzani, G. Pagetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of F-formations," in *British Machine Vision Conference (BMVC)*, 2011.

[39] H. Hung and B. Kröse, "Detecting f-formations as dominant sets," in *International Conference on Multimodal Interfaces (ICMI)*, ser. ICMI '11. New York, NY, USA: ACM, 2011, pp. 231–238.

[40] M. Feldmann, D. Fränken, and W. Koch, "Tracking of extended objects and group targets using random matrices," *IEEE Trans. on Signal Processing*, vol. 59, pp. 1409–1420, 2011.

[41] B. Lau, K. Arras, and W. Burgard, "Multi-model hypothesis group tracking and group size estimation," *International Journal of Social Robotics*, vol. 2, no. 1, pp. 19–30, 2010.

[42] S. J. Mckenna, S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld, "Tracking groups of people," *Computer Vision and Image Understanding*, 2000.

[43] F. Cupillard, F. Brémond, M. Thonnat, I. S. Antipolis, and O. Group, "Tracking groups of people for video surveillance," in *University of Kingston (London)*, 2001.

[44] J. S. Marques, P. M. Jorge, A. J. Abrantes, and J. M. Lemos, "Tracking groups of pedestrians in video sequences," in *IEEE Conference on Computer Vision and Patter Recognition Workshops (CVPR workshops)*, vol. 9, 2003, pp. 101–101.

[45] S. K. Pang, J. Li, and S. Godsill, "Models and algorithms for detection and tracking of coordinated groups," in *Symposium of image and Signal Processing and Analisys*, 2007.

[46] T. Mauthner, M. Donoser, and H. Bischof, "Robust tracking of spatial related components," in *International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[47] L. Bazzani, M. Cristani, and V. Murino, "Collaborative particle filters for group tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2010.

[48] J. Sochman and D. Hogg, "Who knows who - inverting the social force model for finding groups," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 830 –837.

[49] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, p. 4282, 1995. [Online]. Available: doi:10.1103/PhysRevE.51.4282

[50] C. McPhail and R. T. Wohlstein, "Using film to analyze pedestrian behavior," *Sociological Methods & Research*, vol. 10, no. 3, pp. 347–375, 1982.

[51] M. Bierlaire, G. Antonini, and M. Weber, "Behavioral dynamics for pedestrians," in *International Conference on Travel Behavior Research*, 2003.

[52] A. Gning, L. Mihaylova, S. Maskell, S. Pang, and S. Godsill, "Group object structure and state estimation with evolving networks and monte carlo methods," *IEEE Trans. on Signal Processing*, vol. 59, no. 4, pp. 1383 –1396, april 2011.

[53] K. Gilholm, S. Godsill, S. Maskell, and D. Salmond, "Poisson models for extended target and group tracking," in *SPIE Conference: Signal and Data Processing of Small Targets*, 2005.

[54] S. Pellegrini and L. V. Gool, "Tracking with a mixed continuous-discrete conditional random field," *Computer Vision and Image Understanding*, no. 0, pp. –, 2012.

[55] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*. Springer Verlag, 2001.

[56] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[57] N. de Freitas, R. Dearden, F. Hutter, R. Morales-Menendez, J. Mutch, and D. Poole, "Diagnosis by a waiter and a mars explorer," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 455 – 468, mar 2004.

[58] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[59] A. Lo, "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *Annals of Statistics*, vol. 1, no. 12, pp. 351–357, 1984.

[60] D. Blei and M. Jordan, "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[61] J. Sethuraman, "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[62] T. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[63] A. Ahmed and E. Xing, "Dynamic non-parametric mixture models and the recurrent chinese restaurant process with applications to evolutionary clustering," in *SIAM International Conference on Data Mining*, 2008.

[64] H. Ishwaran and L. F. James, "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, vol. 96, pp. 161–173, 2001.

[65] E. T. Hall, *The Hidden Dimension*. Anchor Books, 1966.

[66] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *European Conference on Computer Vision (ECCV)*, 2004, pp. 28–39.

[67] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking multiple and compound visual objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 560–576, 2000.

[68] L. Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, pp. 325–376, December 1992.

[69] M. Isard and A. Blake, "Condensation: Conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998.

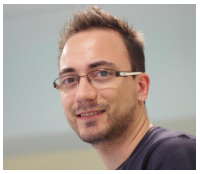
[70] D. Davies and D. Bouldin, "A cluster separation measure," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, no. 2, pp. 224 –227, 1979.

[71] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *International Conference on Computer Vision (ICCV)*, 2009.

[72] K. Smith, D. Gatica-Perez, J.-M. Odobez, and S. Ba, "Evaluating multi-object tracking," in *CVPR*, 2005, p. 36.

[73] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *Journal on Image and Video Processing*, pp. 1–10, Jan. 2008.

[74] 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*. IEEE, 2013, pp. 735–742.



Loris Bazzani is a postdoc at the Dartmouth College. From 2012 to 2013, he was a postdoc at the Istituto Italiano di Tecnologia. He received the BSc and the MSc in intelligent and multimedia systems from the University of Verona in 2006 and 2008 respectively. He obtained the PhD in computer science from the University of Verona in 2012. In 2010, he was visiting student at the University of British Columbia. His research interests include dynamic Bayesian networks, deep learning, kernel methods, large-scale object recognition, person re-identification and tracking.

Matteo Zanotto is a postdoc at the Istituto Italiano di Tecnologia (IIT) in the department of Pattern Analysis and Computer Vision. He received his BSc in Management and Production Engineering (2005) and his MSc in Management, Economics and Industrial Engineering (2007) from the Politecnico di Milano. After working as a Business Process Re-engineering analyst for a Fair-Trade organization, he got his MSc in Artificial Intelligence from the University of Edinburgh (2010) and his PhD (2014) working in IIT on machine learning and computer vision. Among his main research interests there are Bayesian machine learning, especially Bayesian Nonparametric models, unsupervised learning and probabilistic modeling of animal vision.

Marco Cristani (Ph.D.) is assistant professor since 2007 at the Università degli Studi di Verona, Department of Computer Science, where he teaches and does research within the VIsion, Processing and Sound lab (VIPS). He is also affiliate with the Istituto Italiano di Tecnologia, Genova, Italy, where he was Team Leader since 2009-12, now Research Affiliate. His interests are focused on generative modeling and in particular on generative embeddings, with applications on social signal processing and multimedia. Dr. Cristani is co-author of more than 120 papers on important international journals and conferences. He is in the technical program committee of social signaling, pattern recognition conferences and organizer of social signaling/video surveillance. Finally, dr. Cristani is also member of the IEEE, ACM and of the IAPR.

Vittorio Murino received the PhD degree in electronic engineering and computer science from the University of Genova, Italy, in 1993. He is currently a full professor and head of the Pattern Analysis and Computer Vision Department at the Istituto Italiano di Tecnologia, Genova, Italy. After receiving the PhD degree, he was first at the University of Udine and, beginning in 1998, at the University of Verona, where he served as the chairman in the Department of Computer Science from 2001 to 2007. His research interests are in computer vision and machine learning, in particular, probabilistic and statistical techniques for image and video processing, with applications on video surveillance and biomedical image and pattern analysis. He is also a member of the editorial boards of the Pattern Recognition, Pattern Analysis and Applications, and Machine Vision and Applications journals, as well as of the IEEE Trans. on Systems Man, and Cybernetics - Part B: Cybernetics. He is a senior member of the IEEE and a fellow of the IAPR