# Decentralized Particle Filter for Joint Individual-Group Tracking

Loris Bazzani[a,b] Marco Cristani[a,b] Vittorio Murino[a,b]
[a]University of Verona, Strada le Grazie, Verona, Italy
[b]Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Via Morego, Genova, Italy
name.surname@itt.it

## Abstract

*In this paper, we address the task of tracking groups of people in surveillance scenarios. This is a major challenge in computer vision, since groups are structured entities, subjected to repeated split and merge events. Our solution is a joint individual-group tracking framework, inspired by a recent technique dubbed decentralized particle filtering. The proposed strategy factorizes the joint individual-group state space in two dependent subspaces where individuals and groups share the knowledge of the joint individual-group distribution. In practice, we establish a tight relation of mutual support between the modeling of individuals and that of groups, promoting the idea that groups are better tracked if individuals are considered, and viceversa. Extensive experiments on a published and novel dataset validate our intuition, opening up to many future developments.*

## 1. Introduction

Group tracking consists in following tight formations of individuals while they are walking or interacting (Fig. 1). This recent open challenge is important in many respects: in computer vision and signal processing, it may help in locating individual targets in the case of missing measurements [21, 22]; in surveillance, it may reveal social bonds between people, owing to a high-level scene awareness [6, 7] or increase re-identification rates [27].

In general, one of the major difficulties of group tracking lies in the high variability of the group entity: splitting, merging, initialization and deletion are frequent events that characterize the life of a group, and that are usually modeled by heuristic rules, yielding to a scarce generalization. In addition, public tracking benchmarks supplied with serious split and merge episodes are rare in the community, proving the early age of this research topic.

This paper offers an elegant yet effective solution for the group tracking: the idea is to perform, at the same time, tracking of individuals *and* groups, that is, *joint individual-group tracking*. This is made possible by a decentralized policy of filtering [5], that factorizes the joint individual-group state space in two conditionally dependent subspaces, so that it is possible to model 1) the single individuals, and 2) the groups given the knowledge of the individuals. We dubbed our proposal as DEEPER-JIGT: DEcentralizEd Particle filtER for Joint Individual-Group Tracking.

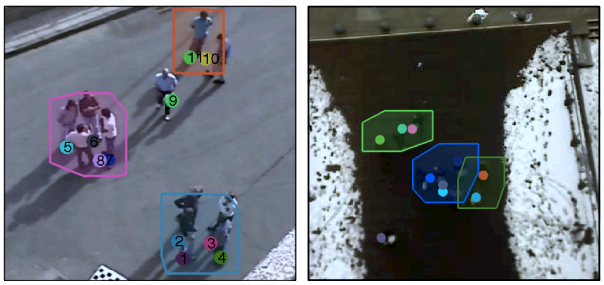


Figure 1: Group tracking results (colored convex hulls) of the proposed Friends meet dataset (first row) and of BIWI dataset (second row) [22].

Many interesting qualities can be ascribed to DEEPER-JIGT. First of all, the absence of heuristics to handle group events: they are all governed by probability distributions whose parameters can be learned from training data. Moreover, DEEPER-JIGT updates the group information in an online fashion, where the tracking history of the individuals is intrinsically exploited by its composite filtering mechanism. This is in contrast with the widely adopted individual-based analysis methods, where groups are estimated by grouping together short individual trajectories (tracklets) collected beforehand, whose length is typically a critical parameter to be tuned [19, 8, 17, 22, 26, 4, 11]. Finally and more important, DEEPER-JIGT allows to understand in a quantitative way how much the modeling of the single targets helps the group tracking *and viceversa*, suggesting that a joint treatment is beneficial for both worlds.

Our proposal has been evaluated on both simulated and real scenarios, providing also a novel benchmark dataset, named *Friends Meet*, where different groups pops out, break out, enter and exit from the scene. To this end, since group tracking evaluation measures do not exist, the existing individual tracking measures [24, 2] have been adapted

here to handle groups.

The rest of the paper is organized as follows. In Sec. 2, a novel taxonomy illustrates the literature on group tracking. The overview of the decentralized particle filter in Sec. 3 introduces our contribution, that is then detailed in Sec. 4. A thorough experimental section is reported in Sec. 5, and, finally, Sec. 6 concludes the paper and envisages the future work.

## 2. Related work

The recent but large literature on group analysis can be partitioned in three categories: 1) the *group-based* class of techniques where groups are treated as genuine atomic entities without the support of individual tracks statistics [25, 10, 12, 15, 16]; 2) the *individual-based* class, where group descriptions are built by associating individuals tracklets that have been calculated beforehand (typically, with a time lag of few seconds) [19, 8, 17, 22, 26, 4, 11]; 3) the *joint individual-group* class, where group tracking and individual tracking are performed simultaneously [21, 18, 1]. Since no extensive essays have been published on this theme yet, interest readers may refer to [15, 11] for good state-of-the-art sections.

The group-based approaches are proposed especially when the scene is highly cluttered so that individual tracking cannot be performed, and the detection of the single targets is unreliable. They assume the groups as nonparametric regions [25], Gaussian-shaped distributions [12, 10], clusters over graphs structures [15], textures [16]. As tracking engines, they employ standard approaches, such as Kalman filtering [12, 10], probability hypothesis density filter [25], multi-hypothesis filtering [15], or particle filtering [16].

In the individual-based category, compact regions are classified as different entities, including groups or persons, exploiting a set of heuristics [19, 8]. In [8], people that stand close for a while are joined into groups through a connection graph built by exploiting heuristics on the moving regions. More principled approaches employ generative modelling [22], discriminative reasoning [26], weighted connection graphs [4] and bottom-up hierarchical clustering [11]. An interesting by-product is presented in [17], where group tracking is employed for facing individual occlusions.

Both the above classes of approaches have in fact drawbacks. The group-based techniques are limited because the individual trajectories are not analyzed, reducing in simplistic models. In the individual-based approaches, the performance is very dependent on the quality of the individual tracklets; more important, groups are seen as mere consequential events of the behavior of the single targets, whereas it is widely known in sociology that groups exert important influence on the acting of the singles.

Joint individual-group techniques deal with individuals and groups at the same time. Many of them maintain the structure of a graph in which connected components correspond to groups of individuals: in [21], stochastic differential equations are embedded in a Markov-Chain Monte Carlo (MCMC) framework, implementing a probabilistic transition model for the group dynamics. The problem of MCMC is that, in its basic form, does not scale efficiently in high-dimensional state spaces. Lately, in [13], a similar framework has been augmented by considering inter-group closeness and intra-group cohesion. In both cases, experiments with few targets are presented. A two-level structure for tracking that uses a physically-based mass-spring model is proposed in [18]: the first level deals with individual tracking, and the second level tracks individuals that are spatially coherent. Similarly in principle, in [1], two processes are involved: the group process considers groups as atomic entities. The individual process captures how individuals move, and revises the group posterior distribution. Both of them do not considers split and merge events.

DEEPER-JIGT lies in this last category, differing from the state of the art in many respects, primarily in the filtering mechanism which was inspired by the Decentralized Particle Filter (DPF) [5]. Moreover, as we will show in the following, DEEPER-JIGT allows to simultaneously deal with merge and split phenomena and with a varying number of individuals and groups. Most important, it allows to understand through quantitative measures the effectiveness of the collaboration of individual and group processes for tracking, promoting the latter category of tracking approaches as the most promising one.

## 3. Decentralized Particle Filter (DPF)

The DPF [5] addresses the classical non-linear discrete-time system

$$\xi_{t+1} = f_t(\xi_t, \eta_t), \quad \mathbf{y}_t = h_t(\xi_t, \eta_t^y) \tag{1}$$

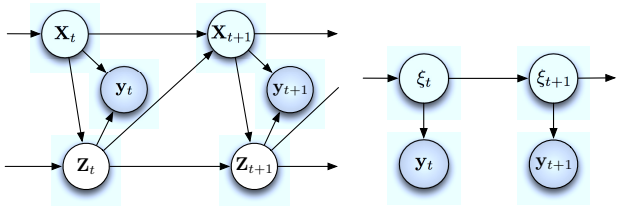
where $\xi_t$ is the state of the system at time $t$, $\mathbf{y}_t$ is the observation or measurement, $\eta_t$ and $\eta_t^y$ are independent non-Gaussian noises, and $f_t$ and $h_t$ are nonlinear functions (Fig. 2(b)). The goal of the DPF is that of recursively estimating the posterior distribution $p(\xi_t|\mathbf{y}_{0:t})$ through a *decomposition* of $\xi_t$ in two (or more) subspaces, *i.e.*, $\xi_t = [\mathbf{X}_t, \mathbf{Z}_t]^T$. Therefore, Eq. 1 can be written as:

$$\mathbf{X}_{t+1} = f_t^x(\mathbf{X}_t, \mathbf{Z}_t, \eta_t^x), \quad \mathbf{Z}_{t+1} = f_t^z(\mathbf{X}_t, \mathbf{Z}_t, \eta_t^z),$$
$$\mathbf{y}_t = h_t(\mathbf{X}_t, \mathbf{Z}_t, \eta_t^y),$$

and the posterior distribution factorizes as:

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

where $\mathbf{X}_{0:t} = (\mathbf{X}_0, \ldots, \mathbf{X}_t)$. In this way, the DPF circumvents both the inefficiency and ineffectiveness of the classical particle filtering when dealing with large sized $\xi_t$s. The




(a) The DPF factorization (b) Joint state space model

Figure 2: Different models for filtering. (a) State decomposition of the joint state $\xi_t = [\mathbf{X}_t, \mathbf{Z}_t]$ with the DPF. (b) Classical particle filtering.

factorization splits the estimation problem in two nested distributions [5]: (i) $p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t})$ and (ii) $p(\mathbf{Z}_t|\mathbf{y}_{0:t}, \mathbf{X}_{0:t})$. Such distributions are analyzed in a serial way, detailed in Alg. 1. The underlying idea is that (i) explains a subspace of the joint space (related to $\mathbf{X}$), and that knowledge is injected into the estimation of $\mathbf{Z}$ in (ii) through the conditional chain rule. More in detail, the DPF performs numerical approximations by importance sampling, explaining both terms of Eq. 2 at time $t$ (steps 1-3), then moving to step $t+1$ (steps 4-7). The distributions highlighted in gray will be explained in the next section. Distributions with subscripts (*e.g.*, $p_{N_z}$) are approximated by samples, and are not described in parametric form.

In Step 1, the standard importance sampling formulation $(Observation \cdot Dynamics)/(Proposal\ distribution)$ is applied for approximating $p(\mathbf{X}_{0:t}|\mathbf{y}_t)$. The difference with the standard framework lies in the term $\mathbf{y}_{0:t-1}$, whose formal presence is motivated by a mathematical derivation discussed in [5] (which is out of the scope of this paper). Intuitively, the conditioning of $\mathbf{y}_{0:t-1}$ injects the knowledge acquired by explaining $\mathbf{y}$ in the $\mathbf{Z}$ subspace at time $t-1$. This highlights the bidirectional relationship of the processes that analyze $\mathbf{X}$ and $\mathbf{Z}$ because, during the same time step, operating on $\mathbf{X}$ helps in better defining $\mathbf{Z}$, and across subsequent time steps, operating on $\mathbf{Z}$ helps $\mathbf{X}$. Step 2 is a classical re-sampling, that regularizes the distributions of the samples (their variance being diminished). Step 3 approximates $p(\mathbf{Z}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ by importance sampling, assuming the dynamics equal to the proposal (so dividing by one). After that, predictions for time $t+1$ are made. As for the previous time step, the $\mathbf{X}$ subspace is first analyzed, sampling particles according to a given dynamics $\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}^{(i)}, \mathbf{y}_{0:t})$. The information encoded in that sample set is plugged into the importance sampling approximation of the posterior $p(\mathbf{Z}_t|\mathbf{X}_{0:t+1}, \mathbf{y}_{0:t})$ (Step 5), yielding to a second resampling step (Step 6) and to the final sampling of $\mathbf{Z}$ at time $t+1$ (Step 7).

In the original paper, the approach was tested with simulations on 2D (4D) points, where $\mathbf{X}$ and $\mathbf{Z}$ lie in two $\mathbb{R}$ ($\mathbb{R}^2$) subspaces. In our case, we are dealing with a much more intriguing and complex problem, where the subspaces have completely different meaning, other than being higher-dimensional. In particular, $\mathbf{X}$ will be the joint state of the individuals, $\mathbf{Z}$ that of the groups. It follows that all the distributions introduced above have been re-designed to fit into the new context.

**Algorithm 1**: The DPF algorithm [5]. INPUT: samples $\{\mathbf{X}_{0:t}^{(i)}\}_{i=1,\dots,N_x}$, samples $\{\mathbf{Z}_{0:t}^{(i,j)}\}_{i=1,\dots,N_x, j=1,\dots,N_z}$. The apices $(i,j)$ mean that for each $i$ particle generated for describing $\mathbf{X}$ we have $N_z$ particles for describing $\mathbf{Z}$. OUTPUT: importance sampling approximations of $\mathbf{X}_{t+1}$, $\mathbf{Z}_{t+1}$.

1. Approximation of $p(\mathbf{X}_{0:t}|\mathbf{y}_t)$ through the importance weights:
$$w_t^{(i)} \propto \frac{p_{N_z}(\mathbf{y}_t|\mathbf{X}_{0:t}^{(i)}, \mathbf{y}_{0:t-1}) p_{N_z}(\mathbf{X}_t^{(i)}|\mathbf{X}_{0:t-1}^{(i)}, \mathbf{y}_{0:t-1})}{\pi(\mathbf{X}_t^{(i)}|\mathbf{X}_{0:t-1}^{(i)}, \mathbf{y}_{0:t-1})}.$$
2. Resample $\{\mathbf{X}_t^{(i)}, \mathbf{Z}_t^{(i,j)}\}$ according to $w_t^{(i)}$.
3. Approximation of $p(\mathbf{Z}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ through the importance weights:
$$\bar{q}_t^{(i,j)} \propto p(\mathbf{y}_t|\mathbf{X}_t^{(i)}, \mathbf{Z}_t^{(i,j)}).$$
4. Generate $\mathbf{X}_{t+1}^{(i)}$ according to $\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}^{(i)}, \mathbf{y}_{0:t})$.
5. Approximation of $p(\mathbf{Z}_t|\mathbf{X}_{0:t+1}, \mathbf{y}_{0:t})$ through the importance weights
$$q_t^{(i,j)} = \bar{q}_t^{(i,j)} p(\mathbf{X}_{t+1}^{(i)}|\mathbf{X}_t^{(i)}, \mathbf{Z}_t^{(i,j)}).$$
6. Resample $\mathbf{Z}_t^{(i,j)}$ according to $q_t^{(i,j)}$.
7. Generation of particles $\mathbf{Z}_{t+1}^{(i,j)}$ according to the proposal $\pi(\mathbf{Z}_{t+1}|\mathbf{X}_{0:t+1}^{(i)}, \mathbf{Z}_t^{(i,j)}, \mathbf{y}_{0:t})$.

## 4. Joint Individual-Group Tracking

Let $\mathbf{X}_t = \{\mathbf{x}_t^k\}_{k=1}^K$ be the joint state of the $K$ individuals at time $t$ and $\mathbf{Z}_t = \{\mathbf{z}_t^k\}_{k=1}^K$ with $\mathbf{z}_t^k \in \{0, 1, \dots, G\}$ be the joint state of the $G$ groups ($K$ and $G$ may vary over time). We define $\mathbf{x}_t^k = (x_t, y_t, \dot{x}_t, \dot{y}_t)$ (individual positions and velocities) and $\mathbf{z}_t^k$ as the group label for the $k$-th individual. As an example, suppose we have 5 individuals and 2 groups at time $t$: with $\mathbf{Z}_t = [1, 1, 2, 2, 0]^T$ we indicate 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 customization of the DPF algorithm for our tracking scenario requires an appropriate redesign of the probability distributions highlighted in gray in Alg. 1. As usual in the particle filtering strategies, distributions may have an analytical form, and/or they can be approximated by particles' sets. In general, one prefers the latter case as this allows one to deal with arbitrarily complex distributions. Analytical functions are usually simpler, typically with Gaussian profiles, but this reduces the expressiveness of the tracking

| Distribution | Analyt. | Approx. |
|---|---|---|
| $\pi(\mathbf{X}_{t+1}\|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ | ✔ | ✔ |
| $p(\mathbf{y}_t\|\mathbf{X}_t, \mathbf{Z}_t)$ | ✔ | ✗ |
| $p(\mathbf{X}_{t+1}\|\mathbf{X}_t, \mathbf{Z}_t)$ | ✔ | ✗ |
| $\pi(\mathbf{Z}_{t+1}\|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t})$ | ✗ | ✔ |

Table 1: Probability $p(\cdot)$ and proposal $\pi(\cdot)$ distributions for DPF. The second and third columns identify which distributions are evaluated and sampled, respectively.

posterior. The situation for DPF is illustrated in Table 1, and such distributions will be defined in the following.

**Individual Proposal** $\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$**.** This distribution models the dynamics of the individuals. Inspired by [20], we adopt the notion of *composite* proposal, incorporating two sources of information:

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

Here, the first part assumes Markovianity between $\mathbf{X}$'s and conditional independence w.r.t. the observation $\mathbf{y}_t$, and adopts a locally linear dynamics with Gaussian noise:

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

where $T$ is the sampling interval and $\eta \sim \mathcal{N}(\mathbf{0}, \Sigma^k)$. 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:

$$\pi(\mathbf{X}_{t+1}|\mathbf{X}_t) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_{t+1}^k | A\mathbf{x}_t^k, \Sigma^k) \tag{3}$$

that is, a multivariate Gaussian distribution with block-diagonal covariance matrix: $\text{diag}(\Sigma^1, \Sigma^2, \ldots, \Sigma^K)$. We can assume $\Sigma^k = \Sigma$ for each $k = 1, \ldots, K$, supposing individuals having usually similar motions.

The second part $\pi_{\text{det}}(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1})$ presumes the presence of a detector. In our case, the distribution is defined as a multivariate Gaussian distribution with the same covariance matrix of Eq. 3 and the positions of the detections associated to each target as means. The parameter $\alpha$ is set once and kept fixed for all the experiments.

**Joint Observation Distribution** $p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{Z}_t)$**.** We adopt a standard template-based technique [3], where the goal is to find the hypothesis that is most similar to a template of the object that is being tracked. To make standard observation models suitable for our framework, we re-write the joint observation distribution as follows:

$$p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{Z}_t) \propto p(\mathbf{Z}_t|\mathbf{y}_t, \mathbf{X}_t)\, p(\mathbf{y}_t|\mathbf{X}_t). \tag{4}$$

In this way, we can model $p(\mathbf{y}_t|\mathbf{X}_t)$ as in standard particle filtering approaches [14]. For simplicity, we define it as:

$$p(\mathbf{y}_t|\mathbf{X}_t) = \prod_{k=1}^{K} p(\mathbf{y}_t|\mathbf{x}_t^k) \propto \prod_{k=1}^{K} exp(-\lambda_{d_y}\, d_y(f(\mathbf{y}_t, \mathbf{x}_t^k), \tau^k))$$

where $d_y$ is a distance between features, $f(\mathbf{y}_t, \mathbf{x}_t^k)$ extracts features from the current bounding box in the image given by $\mathbf{x}_t^k$ and $\tau^k$ is the template of the $k$-th individual[1].

We also assume conditional independence between $\mathbf{Z}_t$ and $\mathbf{y}_t$, *i.e.*, $p(\mathbf{Z}_t|\mathbf{y}_t, \mathbf{X}_t, ) = p(\mathbf{Z}_t|\mathbf{X}_t)$. This term models the likelihood that $\mathbf{Z}_t$ has been generated from $\mathbf{X}_t$. Each group hypothesis $\mathbf{Z}_t^{(i,j)}$ can be seen as a clustering hypothesis of the data $\mathbf{X}_t^{(i)}$. Hence, $p(\mathbf{Z}_t|\mathbf{X}_t)$ can be formulated in terms of *cluster validity* evaluation as follows:

$$p(\mathbf{Z}_t|\mathbf{X}_t) \propto exp(-\lambda_{d_{cl}}\, d_{cl}(\mathbf{Z}_t, \mathbf{X}_t))$$

where $d_{cl}(\mathbf{Z}_t, \mathbf{X}_t)$ is a cluster validity measurement of the hypothesis $\mathbf{Z}_t$ with respect to $\mathbf{X}_t$. Among the different cluster validity measurements, we choose the Davies-Bouldin index [9], because of its simplicity and versatility.

**Joint Individual Distribution** $p(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{Z}_t)$**.** This distribution models the dynamics of the individual taking into account the presence of the group:

$$\mathbf{x}_{t+1}^k = \mathbf{x}_t^k + B\mathbf{g}_t^k + \eta \tag{5}$$

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

$\mathbb{I}(\cdot)$ is the indicator function and $\mathbf{g}_t^k$ is the position and velocity of the group the $k$-th individual belongs to. This term mirrors the fact that individuals in the same group should have similar dynamics. Similarly to Eq. 3, the resulting probability distribution is:

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

**Joint Group Proposal** $\pi(\mathbf{Z}_{t+1}|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t})$**.** The joint group proposal models the dynamics of the groups, and assumes the form

$$\pi(\mathbf{Z}_{t+1}|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t}) = \\ = f(\prod_g \pi(e_{t+1}^g|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t}), \mathbf{Z}_t) \tag{6}$$

$$= f(\prod_g \pi(e_{t+1}^g|\mathbf{X}_{0:t+1}, g_t, g_t', \mathbf{y}_{0:t}), \mathbf{Z}_t) \tag{7}$$

[1]We use Bhattacharyya distance between RGB color histograms and the template is never updated. Note that plenty of more sophisticated techniques could fit our framework.

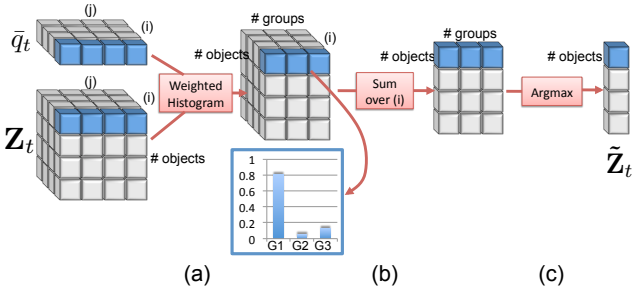



Figure 3: Computation of the state estimate $\tilde{\mathbf{Z}}_t$ that deals with discrete labels.

where the *surrogate* distribution $\pi(e^g_{t+1}|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t})$ in Eq. 6 operates by assigning probabilities on the *events* related to the $g$-th group, *i.e.*, $e^g \in \{\text{Merge}, \text{Split}, \text{None}\}$. In other words, given a group configuration $\mathbf{Z}_t$, whose individuals moved as recorded in $\mathbf{X}_{0:t+1}$, we want to model the probability that a merge or split event occurs, or that the group assignment of each individual remains unchanged. To simplify the modeling, the surrogate is rewritten as in Eq. 7, considering only interactions between a group $g$ and its nearest group $g'$. The deterministic function $f$ translates a selected event in a novel configuration $\mathbf{Z}_{t+1}$, changing the label assignment of $\mathbf{Z}_t$, enlarging or diminishing its size if novel objects (dis)appear. Note that in our approach, a group is an entity formed at least by two individuals.

The distribution $\pi(e^g_{t+1}|\mathbf{X}_{0:t+1}, g_t, {g'}_t, \mathbf{y}_{0:t})$ is offline learned, adopting the multinomial logistic regression. To this end, a set of possible scenarios containing events have been simulated and labelled. We use as features 1) the inter-group distance between $g$ and the nearest group $g'$, considering their positions and sizes ($d_{KL}$, 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 ($d_{\text{intra}}$). 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 $t+1$.

Once the model has been trained, performing inference is straightforward. Given an existing group $g$, $(d_{KL}, d_v, d_{\text{intra}})$ for time $t$ and $t+1$ are computed and fed into the classifier, obtaining an estimate of $\pi(e^g_{t+1}|\mathbf{X}_{0:t+1}, g_t, {g'}_t, \mathbf{y}_{0:t})$. A new event $e^g_{t+1}$ is sampled from that distribution. Note that sampling from it is easy and efficient, because it is discrete and the set of possible events is relatively small. Once the event $e^g_{t+1}$ has been sampled from the proposal distribution, the function $f(\cdot)$ performs the action corresponding to the selected event to generate $\mathbf{Z}_{t+1}$.

In addition, we add a prior over the events in order to reduce the merge between too-large groups. The prior is defined as $\mathcal{N}(|\mathbf{e}^g_{t+1}|; \mu, \sigma)$ where $|\mathbf{e}^g_{t+1}|$ is the size of the $g$-th group after the event $g$ (in the experiments, $\mu = 1$ and $\sigma = 1.5$).

**State Estimate.** In this section, we describe how to estimate the most likely joint state. The joint probability distribution $p(\mathbf{Z}_t, \mathbf{X}_{0:t}|\mathbf{y}_{0:t})$ can be estimated by the DPF once defined each probability distribution in Table 1. The joint state is usually defined as the expected value of the state under a certain distribution, that is, $\mathbf{X}_t = \mathbb{E}_{p(\mathbf{X}_t|\mathbf{y}_{0:t})}$ and $\mathbf{Z}_t = \mathbb{E}_{p(\mathbf{Z}_t|\mathbf{y}_{0:t})}$. Using the empirical approximation given by the DPF, we can easily estimate $\mathbf{X}_t$ as $\tilde{\mathbf{X}}_t = \sum_{i=1}^{N_x} w_t^{(i)} \mathbf{X}_t^{(i)}$.

Since the domain $\mathbf{Z}_t$ is based on discrete labels, the expectation operation cannot be performed. Instead, we compute a distribution over the possible labels as depicted in Fig. 3. Starting from the matrices $\mathbf{Z}_t$ and $\bar{q}_t$, we compute the following distribution for the $k$-th individual as weighted histogram:

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

This gives a similar representation of the sum over $j$ but it considers labels $g$ (step (a) in Fig. 3). Then, each $\text{Wh}^{k,(i,g)}$ 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 $\tilde{\mathbf{Z}}_t$ (step (c) in Fig. 3).

## 5. Experiments

This section shows the potentialities of DEEPER-JIGT in performing joint individual-group tracking on different datasets, while investigating the effects of the mutual support of the group and the individual tracking processes. The structured filtering architecture of DEEPER-JIGT allowed to achieve this goal, by inhibiting conditional dependencies in distributions where mixed terms ($\mathbf{X}$ and $\mathbf{Z}$) do appear.

**Datasets.** The ideal benchmark should handle a scenario where *labelled* groups of people are evolving, appearing and disappearing spontaneously, experiencing split and merge events. This correspond to cocktail party-like situations, *i.e.*, focusing on social areas where people arrive alone or with other people, move from one group to another, stay still while conversing, etc. Nowadays, such a picture is missing, since almost all the existent datasets with labelled groups report different situations, mainly wandering people following a main flow direction (*e.g.*, [23]). In this case, groups are mostly limited to very few people (mostly couples) and the frequency of merge and split is low.

For these reasons, we propose a novel dataset, freely downloadable at `http://goo.gl/cFXCG`, dubbed *Friends Meet*. It is composed by 53 sequences, for a total of 16286 frames. The sequences are partitioned in a synthetic set (28 sequences, 200 frames each), with the aim of stressing tracking strategies in capturing group events, without any complex object representation (simple colored

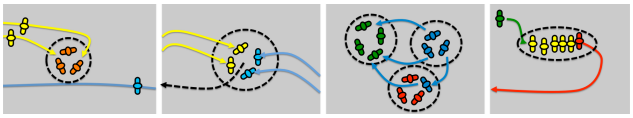


Figure 4: Typical scenarios in the *Friends Meet* dataset: merge and split between groups, queue, and complex situations.

blobs), and a real dataset. In the *synthetic* set, 18 sequences are simple, containing 1-2 events with 4-10 individuals; the other 10 sequences are more challenging, with 10-16 individuals involved in multiple events.

The *real* set focuses on an outdoor area where people usually meet during coffee breaks. This area has been recorded and annotated by an expert for one month. The expert reported the events appeared more frequently, building a screenplay where these events are summarized in order to limit the dataset size. Therefore, the screenplay was played by students and employees, resulting in 15 sequences of different length (between 30 sec. to 1.5 minutes), judged by the expert as sufficiently realistic. In total, the sequences contain from 3 to 11 individuals, and all of them are ground truthed with individual and group information. Some typical scenes are depicted in Fig. 4.

In addition, we provide both quantitative and qualitative results using the BIWI dataset [23], even though it is not well-suited for our method because group events are absent.

**Evaluation Metrics.** The evaluation of DEEPER-JIGT considered the individual tracking and the group tracking results. Unfortunately, while there exists a lot of standard evaluation metrics for individual tracking, there are no widely-accepted measures for group tracking. For this reason, we cast the metrics proposed in [24, 2] to deal with groups.

For individual tracking, we employ Mean Square Error (MSE) over the positions of the individuals, and its standard deviation. The group results have been evaluated by adapting the metrics proposed in [24] for detection (False Positive (FP) and False Negative (FN)) and in [2] for tracking (Multi-Object Tracking Precision (MOTP) and Accuracy (MOTA)). The notion of person bounding box is sobstituted to that of convex hull around the members of groups. Thus, intersection operations among bounding boxes translate naturally in that of convex hulls. We also introduced the Group Detection Success Rate (GDSR) as the detection rate over time of the correctly detected groups. A group is *correct* if at least the 60% of its members are detected [7].

**Results.** The evaluation focused first on the synthetic part of the Friends Meet dataset. For the investigation of the mutual support of the group and the individual tracking processes, we build three variants of DEEPER-JIGT, that is, VAR1, VAR2 and VAR3 .

| | MSE px (std) | 1-FP | 1-FN | GDSR | MOTP px | MOTA |
|---|---|---|---|---|---|---|
| DEEPER-JIGT | **2.18 (4.96)** | 93.74% | **82.94%** | **79.65%** | **16.66** | **57.28%** |
| VAR1 | 2.19 (5.46) | **93.77%** | 81.86% | 78.25% | 17.74 | 55.99% |
| VAR2 | 3.72 (11.81) | 82.11% | 51.61% | 48.09% | 151.03 | 33.53% |
| VAR3 | 2.52 (8.35) | 65.09% | 24.56% | 18.89% | 397.85 | 4.86% |

Table 2: Results on synthetic sequences: 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.

VAR1 assumes $p(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{Z}_t) = p(\mathbf{X}_{t+1}|\mathbf{X}_t)$, inhibiting the contribute of the group in defining the dynamics of the individual, by canceling out the $B\mathbf{g}_t^k$ term of Eq. 5. VAR2 is equal to VAR1, assuming in addition $\pi(\mathbf{Z}_{t+1}|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t}) = \pi(\mathbf{Z}_{t+1}|\mathbf{Z}_t, \mathbf{y}_{0:t})$, that is, suppressing the knowledge of the individual state in promoting events for the group 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 them distributed in a uniform fashion. From DEEPER-JIGT to VAR2, we can notice that the distributions become conditionally independent, and thus sampling is performed indipendently in each state space. Only the observation model links them. Finally, VAR3 is VAR2 with $p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{Z}_t) = p(\mathbf{y}_t|\mathbf{X}_t)$, blocking the contribution of the clustering evaluation, i.e., fixing $p(\mathbf{Z}_t|\mathbf{y}_t, \mathbf{X}_t) = 1$ in Eq. 4. This way, the model judges the individuals, but not how well they fit in groups hypotheses. In practice, this variant separates the individual tracking from the group tracking in two different particle filters.

The results on the synthetic data are summarized in Table 2. The first significant message is clear: the components of DEEPER-JIGT are all needed for reaching the best performances. Moreover, all the performance measures decrease when incrementally pruning away connections between the individuals and the groups (from VAR1 to VAR3).

Actually, the performance of VAR1 tells that in a joint individual-group tracking framework, the individual dynamics should consider the influence that the group exerts on the single person. This helps just a little the group description, while it is uninfluential if we focus on the individual tracking only. We think that this relationship could be exploited more effectively if more advanced group-driven dynamics are injected, *e.g.*, [23, 26, 7].

The performance of VAR2 suggests that the dynamics of a group (intended as the possibility of splitting or merging) cannot be treated as an independent process, and must necessarily be linked to the behavior of the single individuals. This is intuitive, and is beneficial for both individual and group tracking. Even in this case, social grouping mechanisms [23, 26, 7] can boost the performances. The perfor-

a) FM dataset

| | 1-FP | 1-FN | GDSR | MOTP m | MOTA |
|---|---|---|---|---|---|
| DEEPER-JIGT.2 | **97.05%** | **93.82%** | **88.46%** | **0.64** | **71.70%** |
| DEEPER-JIGT | 95.61% | 91.13% | 86.11% | 0.80 | 67.58% |
| VAR3 | 74.77% | 37.72% | 25.92% | 2.80 | 2.73% |

b) BIWI dataset

| | 1-FP | 1-FN | GDSR | MOTP m | MOTA |
|---|---|---|---|---|---|
| DEEPER-JIGT | 53.77% | **78.00%** | **53.59%** | **0.44** | **29.43%** |
| VAR3 | **60.55%** | 51.57% | 29.60% | 1.03 | 9.58% |

Table 3: Group results on a) the FM dataset and b) the BIWI dataset: group detection (columns 2-4) and group tracking (column 5-6). For MOTP (in meters), the lower the better.

mance of VAR3 is the most enlightening: it shows that for modeling groups and individuals a joint treatment is highly recommendable, being the performances of the two separate processes strongly inferior to DEEPER-JIGT.

The second analysis takes into account the real datasets. Since these datasets are very challenging for tracking, due to occlusions and low resolution, a track is re-initialized from the ground truth when the target is lost (distance of $0.6$ meters). The mean re-initialization rate for a target is $3.2\%$ for the real FM dataset. We compare DEEPER-JIGT against VAR3 and DEEPER-JIGT.2. In DEEPER-JIGT.2, we assume that $p(\mathbf{X}_{0:t}|\mathbf{y}_t)$ is completely known, that is, at each time the individual tracker is initialized from the ground truth. In other words, this variant of the algorithm evaluates the method when very low uncertainty on individual tracking is present, thus representing an upper bound on the group performances.

The group tracking accuracies on the FM dataset and the BIWI dataset are summarized in Table 3(a-b). The table highlights the increase of the performance from VAR3 to DEEPER-JIGT. Differently from the synthetic scenario, the false positive rates ($1-$FP) of the different methods are close (Table 3(a)). The low value of $1-$FN for VAR3 mirrors the fact that the method looses the $56\%$ of the groups. The other metrics follow the trend of the results on the synthetic dataset.

Comparing DEEPER-JIGT and DEEPER-JIGT.2 (Table 3(a)), it is interesting to note that if $p(\mathbf{X}_{0:t}|\mathbf{y}_t)$ is known, we obtain very similar results. This means that the uncertainty in the process does not affect very much the joint individual-group tracking. Moreover, Table 3(b) shows that even if the BIWI dataset is harder due to the low resolution and does not contain groups event, DEEPER-JIGT is still able to get reasonable results.

Qualitative results of DEEPER-JIGT on FM dataset (rows 1-3) and BIWI dataset (row 4) are reported in Fig. 5 and the video at `http://youtu.be/J_HDJflQATo`. The figure shows different examples of merge (row 1), initialization and split (row 2), and more complex scene where multiple events occur (row 3). In the sequence `seq_eth` of BIWI dataset (row 4), we noticed that DEEPER-JIGT is able to capture groups of wandering people, even in the case of crowd.

## 6. Conclusions

This study promotes the joint online treatment of individuals and groups in tracking applications. Apart from sociological matters (people may decide to move differently whether they are alone or not), we showed here that this strategy is convenient *quantitatively*. As tracking strategy, we have been inspired from a brand-new filtering mechanism named Decentralized Particle Filter, leading to the design of the DEEPER-JIGT framework. The acronym mirrors its potentially deep customizability, that allows to tweak many filtering mechanisms, defining the dynamics of the group given the individual states and viceversa, how observations are evaluated, etc., as modules of a serial framework. This is indeed a first attempt which proved to deserve further investigation. Next steps will be devoted to ameliorate the dynamics modules, possibly by embedding social force models as individual dynamics, improving how groups are evaluated, for example by importing social signal processing notions. At the same time, the inherent parallelization of the Decentralized Particle Filter, neglected in this paper and ignored in the coding of DEEPER-JIGT, can be taken into account.

## References


[1] L. Bazzani, M. Cristani, and V. Murino. Collaborative particle filters for group tracking. In *IEEE ICIP*, 2010.

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

[3] L. Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24:325–376, December 1992.

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

[5] T. Chen, T. Schon, H. Ohlsson, and L. Ljung. Decentralized particle filter with arbitrary state decomposition. *IEEE Trans. on Signal Processing*, 59(2):465–478, 2011.

[6] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *International Workshop on Visual Surveillance*, pages 1282–1289, 2009.

[7] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, A. D. Bue, D. Tosato, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, 2011.

[8] 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.

[9] D. Davies and D. Bouldin. A cluster separation measure. *IEEE Trans. on PAMI*, (2):224 –227, 1979.

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

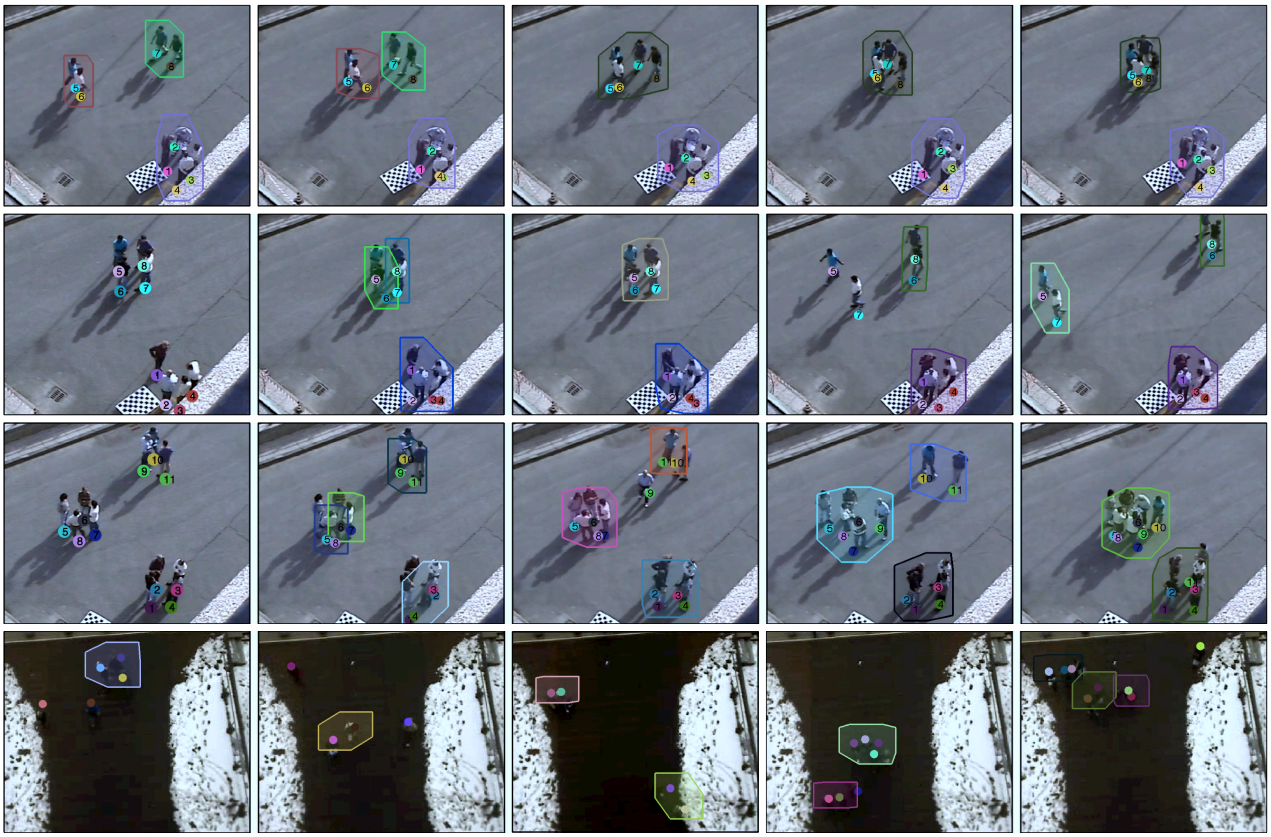


Figure 5: Qualitative results on selected sequence of the FM dataset (first three rows) and BIWI dataset (last row). 1st row: merging between two groups. 2nd row: split event, showing that when people are too far, they are detected as separate persons. 3rd row: a person moves from one group to another, then a merge between two groups occurs.


[11] W. Ge, R. Collins, and R. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Trans. on PAMI*, 99(PrePrints), 2011.

[12] G. Gennari and G. Hager. Probabilistic data association methods in visual tracking of groups. In *CVPR*, 2004.

[13] A. Gning, L. Mihaylova, S. Maskell, S. Pang, and S. Godsill. Group object structure and state estimation with evolving networks and monte carlo methods. *Signal Processing, IEEE Transactions on*, 59(4):1383 –1396, april 2011.

[14] M. Isard and A. Blake. Condensation: Conditional density propagation for visual tracking. *Int. J. of Computer Vision*, 29:5–28, 1998.

[15] B. Lau, K. Arras, and W. Burgard. Multi-model hypothesis group tracking and group size estimation. *I. J. Social Robotics*, 2(1):19–30, 2010.

[16] W.-C. Lin and Y. Liu. A lattice-based MRF model for dynamic near-regular texture tracking. *IEEE Trans. on PAMI*, 29(5):777–792, 2007.

[17] J. S. Marques, P. M. Jorge, A. J. Abrantes, and J. M. Lemos. Tracking groups of pedestrians in video sequences. In *Workshop at CVPR*, volume 9, pages 101–101, 2003.

[18] T. Mauthner, M. Donoser, and H. Bischof. Robust tracking of spatial related components. In *ICPR*, pages 1–4, 2008.

[19] S. J. Mckenna, S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld. Tracking groups of people. *CVIU*, 2000.

[20] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, pages 28–39, 2004.

[21] 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.

[22] S. Pellegrini, A. Ess, and L. V. Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, pages 452–465, 2010.

[23] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009.

[24] K. Smith, D. Gatica-Perez, J.-M. Odobez, and S. Ba. Evaluating multi-object tracking. In *CVPR*, page 36, 2005.

[25] 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 *ICPR*, 2006.

[26] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg. Who are you with and where are you going? In *IEEE CVPR*, 2011.

[27] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.