Supplementary Material: Joint Individual-Group Modeling for Tracking

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

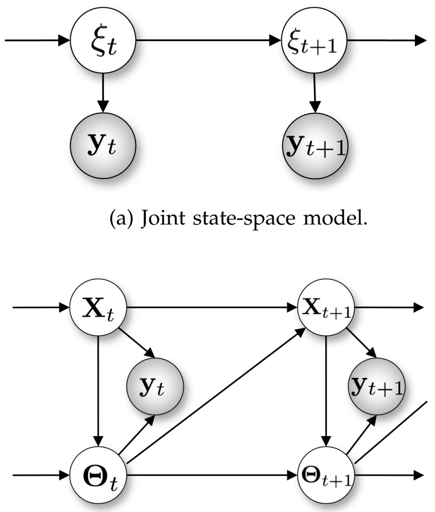
Abstract—In this document, we extend the main manuscript with a detailed descriprion of the formulation of the joint individual-group tracking problem, the derivation of the proposed model and the algorithm. The goal is to give to the reader a better insight on the proposed solution and the type of interactions that exist between the two subspaces $\mathbf{\Theta}$ and $\mathbf{X}$. In addition, the details of the joint group proposal used in the DEEPER-JIGT are discussed to make it clearer. Finally, we perform an extended analysis of the results aimed at understanding which scenarios are better for the DEEPER-JIGT and which for the DP2-JIGT. We also attached to this document a video showing the qualitative results of the presented model on different datasets.

1 Derivations and Details of the Model

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 noises, and f_t and h_t are nonlinear unknown functions. Eq. 1 practically leads to the conditional probabilities defined by the link from ξ_t to ξ_{t+1} and the link from ξ_t to $\mathbf{y}_t$ in Fig. 1a.



(a) Joint state-space model.

(b) The proposed model.

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

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 $\mathbf{\Theta}_t$.

The system of Eq. 1 is thus rewritten 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 DEEPER-JIGT and the DP2-JIGT (Fig. 1b) are instances of this general formulation. In both models, the state of the individuals $\mathbf{X}_{t+1}$ depends on its previous state $\mathbf{X}_t$ and the previous state of the groups $\mathbf{\Theta}_t$, and the state of the groups $\mathbf{\Theta}_{t+1}$ depends on the current state of the individuals $\mathbf{X}_{t+1}$ and the previous state of the groups $\mathbf{\Theta}_t$. These relations encode the interdependence between the two subspaces and are built into the model through defining appropriate conditional probability densities. 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.

Following the line of reasoning of [1], we show in the following how 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. 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 $\mathbf{X}_t$ is the individual state variable, $\mathbf{\Theta}_t$ is the group state variable, and $\mathbf{y}_{0:t} = (\mathbf{y}_0, \ldots, \mathbf{y}_t)$ 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.

- *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 is also with the Department of Computer Science, University of Verona, Verona, Italy.*

1.1 First subproblem: $p(\boldsymbol{\Theta}_t|\mathbf{X}_t,\mathbf{y}_{0:t})$

The first term of Eq. 2 becomes:

$$\begin{aligned} p(\boldsymbol{\Theta}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t}) &\propto p(\boldsymbol{\Theta}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})p(\mathbf{y}_t|\mathbf{X}_{0:t},\boldsymbol{\Theta}_t) \quad (3)\\ &= p(\boldsymbol{\Theta}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})p(\mathbf{y}_t|\mathbf{X}_t,\boldsymbol{\Theta}_t) \quad (4)\end{aligned}$$

where the independence assumption $p(\mathbf{y}_t|\mathbf{X}_{0:t},\boldsymbol{\Theta}_t) = p(\mathbf{y}_t|\mathbf{X}_t,\boldsymbol{\Theta}_t)$ is shown in the model of Fig. 1b.

The first term of Eq. 4 has to be expressed in term of $\boldsymbol{\Theta}_{t-1}$ to obtain the classic recursion of the particle filter:

$$\begin{aligned} &p(\boldsymbol{\Theta}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1}) =\\ &= \int p(\boldsymbol{\Theta}_t,\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_{t-1}\\ &= \int p(\boldsymbol{\Theta}_t|\boldsymbol{\Theta}_{t-1},\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_{t-1}\\ &= \int p(\boldsymbol{\Theta}_t|\boldsymbol{\Theta}_{t-1},\mathbf{X}_t)p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_{t-1} \quad (5)\end{aligned}$$

where in the last equivalence we applied the conditional independence assumptions derived directly from the model of Fig. 1b to the first term.

The second term of Eq. 5 is factorized as:

$$\begin{aligned} &p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1}) \propto\\ &\quad \propto p(\mathbf{X}_t|\mathbf{X}_{0:t-1},\boldsymbol{\Theta}_{t-1},\mathbf{y}_{0:t-1})p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})\\ &\quad = p(\mathbf{X}_t|\mathbf{X}_{t-1},\boldsymbol{\Theta}_{t-1})p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1}) \quad (6)\end{aligned}$$

where we applied the conditional independence assumptions derived from the model of Fig. 1b. Now, it is possible to see the recursion in the inference process: $p(\boldsymbol{\Theta}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t})$ in Eq. 4 is estimated given $p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})$ in Eq. 6. Therefore, sequential importance sampling can be applied to solve the integral of Eq. 5, where the proposal distribution is $\pi(\boldsymbol{\Theta}_t|\boldsymbol{\Theta}_{t-1},\mathbf{X}_t) = p(\boldsymbol{\Theta}_t|\boldsymbol{\Theta}_{t-1},\mathbf{X}_t)$ in our experiments.

1.2 Second subproblem: $p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t})$

The second term of Eq. 2 is factorized as follows:

$$\begin{aligned} &p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t}) =\\ &= p(\mathbf{X}_t|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})p(\mathbf{X}_{0:t-1}|\mathbf{y}_{0:t-1})\\ &\propto p(\mathbf{X}_t|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})p(\mathbf{y}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})p(\mathbf{X}_{0:t-1}|\mathbf{y}_{0:t-1}). \quad (7)\end{aligned}$$

The recursion is straightforward because $p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t})$ depends directly from $p(\mathbf{X}_{0:t-1}|\mathbf{y}_{0:t-1})$. However, the first two term of the equation should be further expanded to introduce the dependency with $\boldsymbol{\Theta}$.

The first term is re-written as follows:

$$\begin{aligned} &p(\mathbf{X}_t|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})\\ &= \int p(\mathbf{X}_t,\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_{t-1}\\ &= \int p(\mathbf{X}_t|\boldsymbol{\Theta}_{t-1},\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_{t-1}\\ &= \int p(\mathbf{X}_t|\boldsymbol{\Theta}_{t-1},\mathbf{X}_{t-1})p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_{t-1} \quad (8)\end{aligned}$$

where we applied the conditional independence assumptions derived from the model of Fig. 1b. This integral is solved through sequential importance sampling given $p(\boldsymbol{\Theta}_{t-1}|\mathbf{X}_{0:t-1},\mathbf{y}_{0:t-1})$.

The second term of Eq. 7 becomes:

$$\begin{aligned} &p(\mathbf{y}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1}) =\\ &\quad = \int p(\mathbf{y}_t,\boldsymbol{\Theta}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_t\\ &\quad = \int p(\mathbf{y}_t|\boldsymbol{\Theta}_t,\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})p(\boldsymbol{\Theta}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_t\\ &\quad = \int p(\mathbf{y}_t|\boldsymbol{\Theta}_t,\mathbf{X}_t)p(\boldsymbol{\Theta}_t|\mathbf{X}_{0:t},\mathbf{y}_{0:t-1})d\boldsymbol{\Theta}_t \quad (9)\end{aligned}$$

where in the last equivalence we applied the conditional independence assumptions derived directly from the model of Fig. 1b to the first term. The first term of Eq. 9 is the observation model, while the second term is computed as in Eq. 5.

2 Joint Group Proposal for the DEEPER-JIGT

In this section, we present in details the approach to learn the joint group proposal followed in [2].

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 designed surrogate distribution is easier to sample from than the original proposal. The joint group proposal for the DEEPER-JIGT is defined as:

$$\begin{aligned} &\pi({}^F\boldsymbol{\Theta}_{t+1}|\mathbf{X}_{t+1},{}^F\boldsymbol{\Theta}_t) =\\ &\qquad = f\Big(\prod_g \pi(e^g_{t+1}|\mathbf{X}_{0:t+1},{}^F\boldsymbol{\Theta}_t),{}^F\boldsymbol{\Theta}_t\Big) \quad (10)\\ &\qquad = f\Big(\prod_{g=1}^{G} \pi(e^g_{t+1}|\mathbf{X}_{t+1},g_t,g'_t),{}^F\boldsymbol{\Theta}_t\Big) \quad (11)\end{aligned}$$

where the *surrogate* distribution $\pi(e^g_{t+1}|\mathbf{X}_{0:t+1},{}^F\boldsymbol{\Theta}_t)$ in Eq. 10 operates by assigning probabilities on the *events* related to the g-th group, *i.e.*, $e^g \in \{\mathrm{Merge},\mathrm{Split},\mathrm{None}\}$. In other words, given a group configuration ${}^F\boldsymbol{\Theta}_t$ and an individual configuration $\mathbf{X}_{t+1}$, 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 and make the problem tractable, the surrogate is rewritten as in Eq. 11, considering only interactions between a group g and its nearest group g'. The deterministic function f translates a selected event in a novel configuration ${}^F\boldsymbol{\Theta}_{t+1}$. This is done by changing the label assignment of ${}^F\boldsymbol{\Theta}_t$ and appropriately modifying its size when groups appear or disappear as a consequence of the split and merge events. Note that in our approach, a group is an entity formed by at least two individuals.

The distribution $\pi(e^g_{t+1}|\mathbf{X}_{0:t+1},g_t,{g'}_t)$ is learned offline through a multinomial logistic regression. In order to obtain training data, we created a naive simulator to generate a set of possible videos/scenarios containing events. There are two reasons for using a simulator: 1) it is straightforward to obtain annotations of the events, and 2) it is usually hard to obtain a significant number of examples of merge and split events from real videos.

Given the simulated scenarios, groups were modelled as Gaussian distributions over observed points (individuals). The following features were extracted and used in the learning stage: 1) inter-group distance between g and the nearest group g', considering their position and size (d_{KL}, symmetrized Kullback-Leibler divergence between Gaussians), 2) inter-group difference between velocities (d_v, Euclidean distance), and 3) the intra-group variance between the positions of the individuals in the g-th group (d_{intra}). The input of the multinomial logistic regression is obtained by concatenating $(d_{KL}, d_v, d_{\text{intra}})$ for time t and $t+1$, resulting in a 6-dimensional vector.

Once the model has been trained, performing inference given a novel feature vector 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 the probability of observing a group split, a merge with the closest group g' or no event. We use this discrete probability as an estimate of $\pi(e^g_{t+1}|\mathbf{X}_{0:t+1}, g_t, {g'}_t)$. A new event e^g_{t+1} is sampled from this distribution. Note that sampling from it is efficient because it is a Discrete distribution 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 ${}^F\mathbf{\Theta}_{t+1}$.

3 Deeper Analysis of the Results

In this section, we perform a deeper analysis of the DEEPER-JIGT or the DP2-JIGT proposed in the main paper on the FM dataset. Fig. 2 shows examples of both synthetic and real data from the dataset.

The aim of this analysis is highlighting in which application scenarios the DEEPER-JIGT or DP2-JIGT should be preferred. To evaluate this, we divided the synthetic and real FM dataset in 5 scenarios:

- *opposite*: individuals and/or groups going in opposite directions but without merging;
- *merge*: individuals and/or groups merge together;
- *split*: groups split in smaller groups or individuals;
- *multiple events*: complex scenario where groups and individuals take part in multiple split and merge;
- *queueing*: individuals are in a queue.

As we discussed in the main paper, our model deals with self-organizing groups, and not with temporary group-like formations induced by external forces. For the sake of completeness, we decided to include the analysis of the queues anyway in this Supplementary Material.

Table 1 shows how the different application scenarios are represented in the Friends Meet datasets.

Table 2 reports the results on the Friends Meet synthetic dataset. The evaluation is performed with the statistics introduced in the main paper. The results of the same analysis for the real FM dataset are reported in Table 3.

Scenario	**FM synth.**	**FM real**
opposite	5	3
merge	5	4
split	5	3
multiple events	10	3
queuing	3	2

TABLE 1: Subdivison of videos of the Friends Meet synthetic (FM synth.) and real (FM real) datasets in the five application scenarios.

From the results in Tables 2 and 3 the following can be highlighted:

- When no events occur (rows 2-3), the DP2-JIGT performs very well.
- When some events occur (rows 4-7), the DP2-JIGT outperforms the DEEPER-JIGT in certain statistics, but it does not perform as well in others (e.g., GDSR row 4 column 6 of Table 2). This is due to the fact that when an event occurs, the online learning method of the DP2-JIGT needs more frames than the offline learning method of the DEEPER-JIGT to learn the new scenario. Once enough frames are observed, though, the online learning strategy is more accurate.
- In case of multiple events (row 8-9), the DEEPER-JIGT has some difficulties to deal with more complex cases with multiple targets and multiple split and merge, because the offline-trainined model is not able to generalize to more complex scenarios, while the online learning method is able to adapt itself to the situation.
- The last 2 rows contain the queue class (excluded on the main paper), where one can notice that the DP2-JIGT have high false positive rate and thus high MOTA. This is the consequence of the social threshold, that in the case of queues keeps oversegmenting the queues in sub-groups.

Some qualitative results that compares the DEEPER-JIGT and the DP2-JIGT are reported in Fig. 3 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 (*e.g.*, S05 $t = 6$). 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 $t = 366$ in Fig. 3), while the DP2-JIGT uses the social constraint to avoid it.

The advantage of the DP2-JIGT over the DEEPER-JIGT is in general due to the fact that the sequences are long enough to online learn the way groups evolve, bringing the DP2-JIGT to generate better hypotheses that have to be evaluated. On the other hand, when the individuals

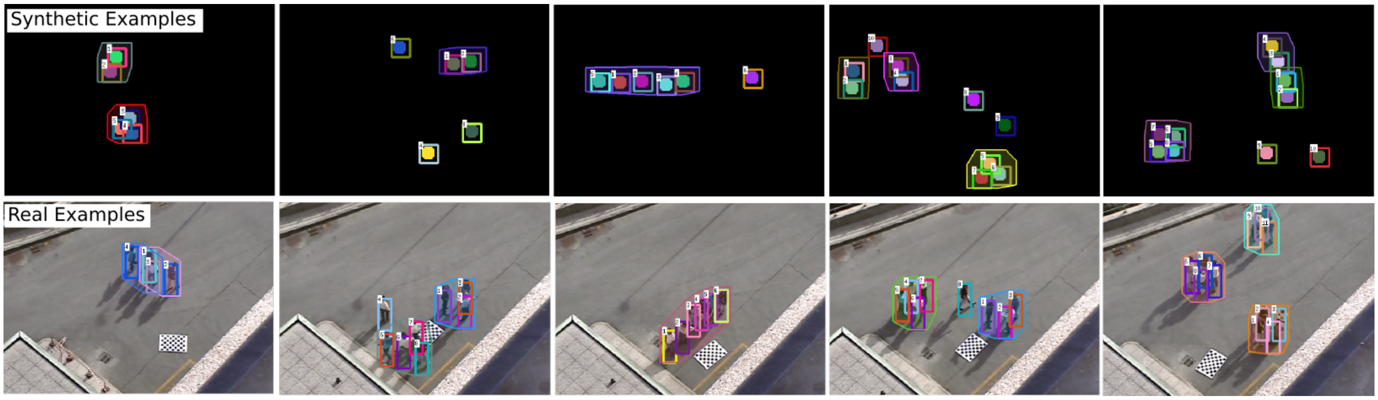


Fig. 2: Some examples of the scenarios contained in the Friends Meet dataset (synthetic and real videos) with the ground truth of individuals and groups.

TABLE 2: Results on the synthetic FM dataset on the different cluster of videos.

Video Cluster	Method	MSE [px] (std)	1-FP	1-FN	GDSR	MOTP [px]	MOTA
Opposite	DP2-JIGT	**0.85** (**0.69**)	**100.00**%	**100.00**%	**100.00**%	**0.82**	**100.00**%
	DEEPER-JIGT	1.01 (0.82)	99.85%	82.10%	81.60%	13.51	65.65%
Merge	DP2-JIGT	**1.24** (**1.16**)	**99.70**%	**89.90**%	75.90%	**2.75**	**82.20**%
	DEEPER-JIGT	1.39 (1.24)	94.50%	89.00%	**82.60**%	9.36	65.09%
Split	DP2-JIGT	**1.21** (**1.28**)	88.10%	**86.05**%	**84.00**%	**4.28**	**67.52**%
	DEEPER-JIGT	1.39 (1.49)	**92.95**%	82.30%	81.85%	6.57	60.78%
Multiple events	DP2-JIGT	**2.72** (**10.34**)	**91.05**%	**90.22**%	**87.33**%	37.88	**54.06**%
	DEEPER-JIGT	3.80 (11.78)	89.16%	75.83%	72.43%	**30.69**	37.82%
Queueing	DP2-JIGT	**1.11** (**1.11**)	56.50%	**99.00**%	**93.17**%	8.46	46.67%
	DEEPER-JIGT	1.39 (1.16)	**98.83**%	**99.00**%	91.83%	**4.14**	**89.33**%

stay on the scene for a limited period of time (*e.g.*, the BIWI sequence "eth"), the DP2-JIGT is able to perform group tracking when the scenario is not much crowded. In this situation, the DEEPER-JIGT is preferred (see last two columns, last two rows in Fig. 3).

References

[1] 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.

[2] 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.

TABLE 3: Results on the real FM dataset on the different cluster of videos.

Video Cluster	Method	1-FP	1-FN	GDSR	MOTP [m]	MOTA
Opposite	DP2-JIGT	**99.61%**	**98.56%**	**96.87%**	**1.01**	**72.65%**
	DEEPER-JIGT	94.28%	72.23%	71.32%	1.54	40.00%
Merge	DP2-JIGT	**98.47%**	**99.00%**	**97.66%**	0.77	**79.96%**
	DEEPER-JIGT	97.55%	96.82%	93.20%	**0.42**	79.04%
Split	DP2-JIGT	**96.90%**	**95.03%**	**91.28%**	0.96	72.53%
	DEEPER-JIGT	94.79%	93.74%	87.36%	**0.58**	**77.05%**
Multiple events	DP2-JIGT	**96.04%**	**97.10%**	**91.79%**	**0.99**	**68.22%**
	DEEPER-JIGT	95.65%	94.91%	88.77%	1.11	60.00%
Queueing	DP2-JIGT	68.31%	96.92%	85.43%	0.58	31.83%
	DEEPER-JIGT	**94.89%**	**98.48%**	**88.22%**	**0.29**	**83.24%**

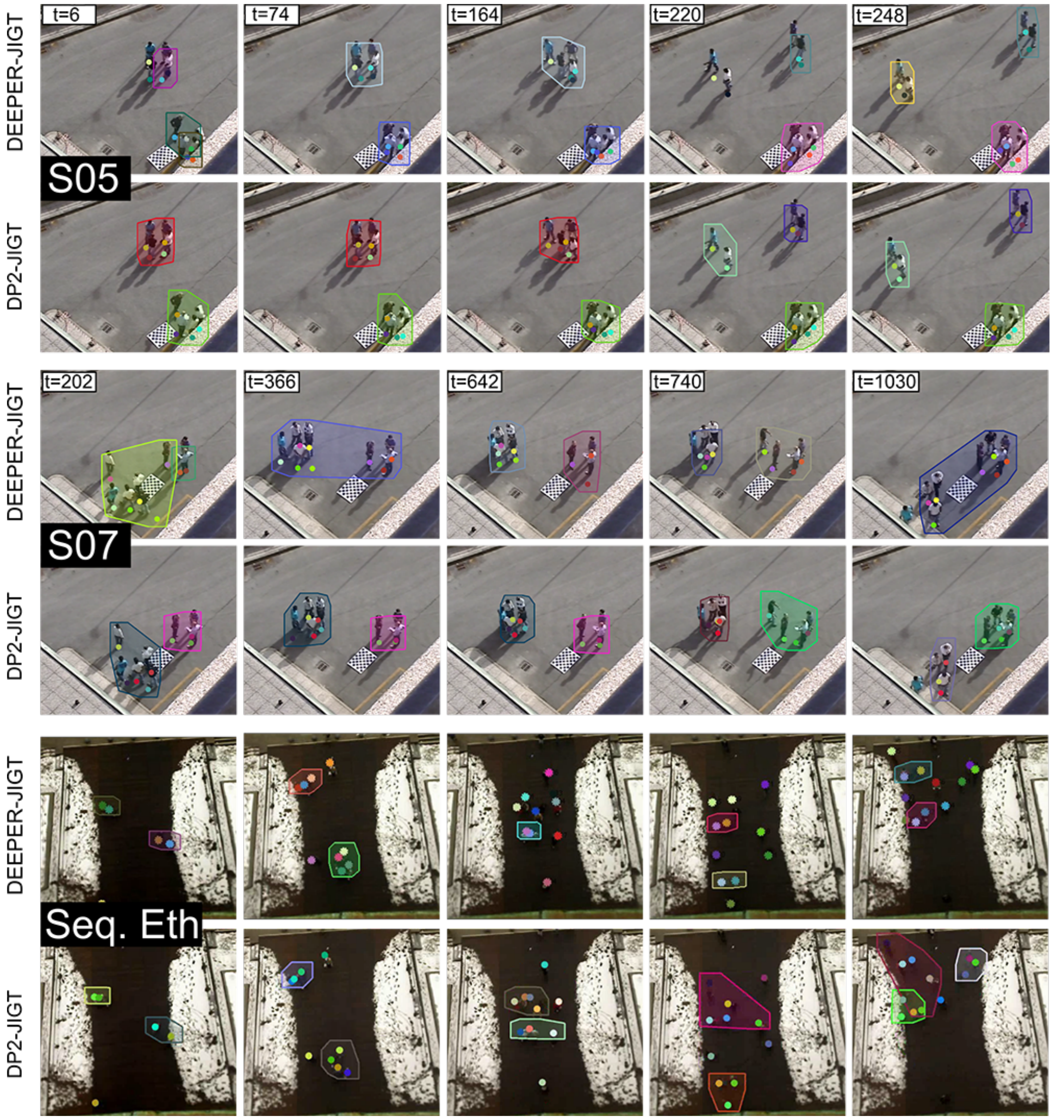


Fig. 3: Qualitative results on sequences of the FM and BIWI datasets comparing the DEEPER-JIGT and the DP2-JIGT.