COLLABORATIVE PARTICLE FILTERS FOR GROUP TRACKING

Loris Bazzani[1], Marco Cristani[1,2], and Vittorio Murino[1,2]

[1] Dipartimento di Informatica, University of Verona - Strada Le Grazie, 15 37134, Verona, Italy
[2] Istituto Italiano di Tecnologia (IIT) - Via Morego 30, 16163, Genova, Italy

ABSTRACT

Tracking groups of people is a highly informative task in surveillance, and it represents a still open and little explored issue. In this paper, we propose a brand new framework for group tracking, that consists in two separate particle filters, one focusing on groups as atomic entities (the *multi-group tracker*), and the other modeling each individual separately (the *multi-object tracker*). The latter helps the multi-group tracker in better defining the nature of a group, evaluating the membership of each individual with respect to different groups, and allowing a robust management of the occlusions. The coupling of the two processes is theoretically founded due to the revision of the posterior distribution of the multi-group tracker with the statistics accumulated by the multi-object tracker. Experimental comparative results certify the goodness of the proposed technique.

Index Terms— Group Tracking, Multi-Target Tracking, Particle Filtering.

1. INTRODUCTION

Group tracking (GT) is of high interest for video surveillance purposes as it allows fine scenario descriptions, addressing choral actions that may lead to important threats, and highlighting social bounds among individuals. At the same time, it represents a challenging issue, since a group of people is a highly structured entity whose dynamics is complex, and whose appearance is erratic, due to intra- and inter-group occlusions phenomena.

There have been few recent attempts to deal with GT problem. In [1], a foreground segmentation method classifies the moving regions in people and groups. In [2], a foreground subtraction-based method models the object paths using a Bayesian network. A set of empirical rules detect the groups. However, in those methods intra- and inter-group dynamics are not considered. A set of empirical merging and splitting rules embedded into a Kalman filter are proposed in [3] to define the groups. However, the Kalman filter is not able to deal with non-linear dynamics. In [4], a deterministic mass-spring model interprets the result of a multi-object tracker, joining objects sharing a common behavior. In [5], a lattice-based Markov Random Field combined to a particle filter tracks groups as near-regular textures. A method that tracks a group of highly correlated targets by employing an Markov Chain Monte Carlo particle filter is proposed in [6]. However, the last two approaches deal with very constrained intra-group dynamics because they assume a strong correlation among the targets.

In this paper, we propose a novel way to track groups, namely Collaborative Particle Filters (Co-PF). The underlying idea consists in designing two tracking processes observing a scene under

two different perspectives: a low-level, *multi-object tracker* (MOT) performs tracking of multiple individuals, separately; a high-level, *multi-group tracker* (MGT) focuses on groups, and uses the knowledge acquired by the MOT to refine its estimations. Each process consists in a Hybrid-Joint Separable (HJS) filter [7], that permits to track multiple entities dealing with occlusions in a very effective way.

The input given by the MOT flows to the MGT in a principled way, *i.e.*, revising the MGT posterior distribution by marginalizing over the MOT's state space. In this way, the MGT posterior is peaked around group configurations formed by trusted individual estimates. In practice, our framework permits to: i) track *multiple* groups; ii) deal with intra- and iii) inter-group occlusions in a 3D calibrated context. The latter two conditions have never been taken into account jointly, and define a brand new operating context, where we put a solid possible solution. Synthetic and real experiments validate our intuition and encourage further developments for Co-PF.

The rest of the paper is organized as follows. In Sec. 2, we describe the mathematical background. In Sec.3, the proposed method is described, and related experiments are presented in Sec.4. Finally, Sec. 5 concludes the paper, indicating the future work for this promising approach.

2. MATHEMATICAL BACKGROUND

From a Bayesian perspective, the single object tracking problem aims at recursively calculating the posterior distribution $p(x_t|z_{1:t})$ by exploiting the Chapman-Kolmogorov equation, where x_t is the current state of the target, z_t is the current measurement, and $x_{1:t}$ and $z_{1:t}$ are the states and the measurements up to time t, respectively. In formulae:

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|z_{1:t-1})dx_{t-1} \quad (1)$$

The equation is fully specified by an initial distribution $p(x_0|z_0) = p(x_0)$, the dynamical model $p(x_t|x_{t-1})$, and the observation model $p(z_t|x_t)$. Particle filtering (PF) approximates the posterior distribution by a set of N weighted particles, *i.e.*, $\{(x_t^{(n)}, w_t^{(n)})\}_{n=1}^N$; a large weight $w_t^{(n)}$ mirrors a state $x_t^{(n)}$ with high posterior probability. In this way, the integral in Eq. 1 has not to be analytically solved, and, instead, the posterior at time $t-1$ is sampled, defining a set of state hypotheses (the particles) that evolve according to the dynamical model $p(x_t|x_{t-1})$ (the prediction step), and which is evaluated via $p(z_t|x_t)$ (the observation step).

HJS filter [7] is an extension of the PF for multiple targets. Defining $\mathbf{x}_t = \{x_t^1, x_t^2, \ldots, x_t^K\}$ the joint state (the ensemble of all individual targets), HJS adopts the approximation $p(\mathbf{x}_t|z_{1:t}) \approx \prod_k p(x_t^k|z_{1:t})$, that is, the joint posterior could be approximated via the product of its marginal components (k indexes the individual
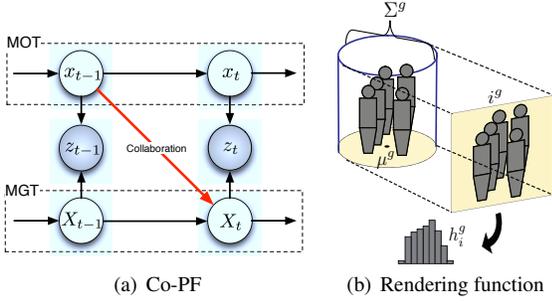
(a) Co-PF   (b) Rendering function

Fig. 1. Collaborative PF idea and group rendering.

targets). The dynamics and the observation models of HJS are expressed as follows:

$$p(x_t^k|x_{t-1}^k) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}^{\neg k}|z_{1:t-1})d\mathbf{x}_{t-1:t}^{\neg k} \quad (2)$$

$$p(z_t|x_t^k) = \int p(z_t|\mathbf{x}_t)p(\mathbf{x}_t^{\neg k}|z_{1:t-1})d\mathbf{x}_t^{\neg k} \quad (3)$$

where the apex $\neg k$ addresses all the targets but the kth. These equations encode an intuitive strategy, i.e., that both the dynamics and the observation phases of the kth target lie upon the consideration of a joint dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1}) \approx p(\mathbf{x}_t) \prod_k q(x_t^k|x_{t-1}^k)$ and observation model $p(z_t|\mathbf{x}_t)$. The joint dynamical model, through the prior $p(\mathbf{x}_t)$, avoids that multiple targets with single motion described by $q(x_t^k|x_{t-1}^k)$ collapse in a single location, and the joint observation model considers that the visual appearance of a single target may be occluded by another object, acting as a z-buffer. The two models are weighted by posterior distributions that essentially promote trusted joint objects configurations (not considering the kth object). For more details, readers may refer to [7].

3. COLLABORATIVE PARTICLE FILTER

Our framework is sketched in Fig. 1(a): the MOT tracks the individuals in the scene, whereas the MGT tracks groups of individuals. Both the processes share the same observations, $\{z_t\}$, and this highlights our key intuition: the two processes evaluate the scene under two different points of view.

The MOT process is modeled by an HJS filter [7]. Each individual state is modeled as an elliptical shape on the ground plane, i.e., $x_t^k = \langle \mu^k, \Sigma^k \rangle$, where μ^k is the position of the individual on the ground plane[1], Σ^k is a covariance that measures the occupancy of the body projected on the ground plane (see [7] for more details).

The MGT process customizes the HJS filter for dealing with groups, and incorporates a fusion component, accepting information from the MOT. We denote the gth group as $X^g = \langle \mu^g, \Sigma^g \rangle$, where μ^g is the 2D position on the floor of the centroid of the gth group and Σ^g is the covariance matrix that approximates the projection of its shape on the floor. The choice of an ellipse for modeling the floor projection of a group is motivated from a psycho-sociological point of view, exploiting proxemics notions that describe a group as a compact closed entity [8]. The posterior of the MGT of the gth group follows the Bayesian recipe (Eq. 1), so that

$$p(X_t^g|z_{1:t}) \propto p(z_t|X_t^g) \int p(X_t^g|X_{t-1}^g) \, p(X_{t-1}^g|z_{1:t-1}) \, dX_{t-1}^g. \quad (4)$$

The dynamical model $p(X_t^g|X_{t-1}^g)$ is derived as in Eq. 2, where the joint dynamical model $p(\mathbf{X}_t|\mathbf{X}_{t-1}) \approx p(\mathbf{X}_t) \prod_g q(X_t^g|X_{t-1}^g)$ has $\mathbf{X}_t = \{X_t^1, X_t^2, \ldots, X_t^G\}$, with G the number of groups in the scene. In this case, the function $q(X_t^g|X_{t-1}^g)$ is modeled by considering the nature of $X_t^g = \langle \mu^g, \Sigma^g \rangle$. For the centroid μ^g, we assume a linear motion, perturbed by white noise with parameter σ_μ. The dynamics of the covariance matrix Σ^g is defined by a perturbation of its principal axes, i.e., by varying its eigenvalues $\{\lambda_i\}_{i=1,2}$ and eigenvectors $\{\mathbf{v}_i\}_{i=1,2}$. In particular, we rotate the principal axes by an angle θ, by modifying the eigenvectors:

$$V' = [R(\mathcal{N}(\theta, \sigma_\theta))\,\mathbf{v}_1,\, R(\mathcal{N}(\theta, \sigma_\theta))\,\mathbf{v}_2] \quad (5)$$

and then, we vary the amplitude of the principal axes by modifying the eigenvalues as follows:

$$\Lambda' = \begin{bmatrix} \mathcal{N}(\lambda_1, \sigma_\lambda) & 0 \\ 0 & \mathcal{N}(\lambda_2, \sigma_\lambda) \end{bmatrix} \quad (6)$$

where $R(\cdot)$ is a rotation matrix and σ_θ and σ_λ are user-defined noise variance values. The matrixes V' and Λ' are then used to recompose the new hypothesis $\Sigma' = V'\Lambda'V'^T$, that will represent a new perturbed elliptical shape.

The dynamics prior $p(\mathbf{X}_t^g)$ implements an exclusion principle [7] that cancels out some inconsistent particles. In other words, two groups' hypotheses, that are close and partially overlapped, will be rejected. We employ a Markov Random Field learned via Belief Propagation, where each node represents a group hypothesis and the weight of each link is computed considering the overlapping area of the two hypotheses.

The (single) observation model $p(z_t|X_t^g)$ is derived from Eq.3, where we have $p(z_t|\mathbf{X}_t)$ as joint observation model. In order to easily evaluate an observation z_t, we employ a *rendering function* that maps a state in a convenient feature space[2]. The idea is depicted in Fig. 1(b): when a new group is detected at time t in the scene [3], its centroid μ^g and occupancy area Σ^g are robustly estimated, forming the initial state X_t^g. The rendering function builds a volume of height 1.80m upon the area Σ^g, in order to surround the people of the group. From this volume, the projection i^g (namely, the model of X_t^g) on the image plane is evaluated, and finally, the histogram h_i^g is computed. This function permits to estimate novel state hypotheses $X_t'^g$: given its components $\langle \mu'^g, \Sigma'^g \rangle$, the rendering function takes the model i^g deforming it opportunely (by a re-scaling, considering the μ'^g, and by a shearing, taking into account the deformation resulted by the perturbation of the covariance matrix Σ'^g). This brings to a novel $h_i'^g$, which is compared with the observation estimated directly from the scene by the rendering function applied to $\langle \mu'^g, \Sigma'^g \rangle$. We use the Bhattacharyya distance as similarity measurement.

The joint observation model $p(z_t|\mathbf{X}_t)$ considers all the groups present in a scene, considering what part of the group X_t^g is seen by taking into account the remaining groups $\mathbf{X}_t^{\neg g}$. This encodes at the same time pro and cons of the observation model. Actually, we assume a group as a rigid solid shape (the model i^g), and this permits to model inter-group occlusions, but it does not model intra-group occlusions (i.e., persons of a group that mutually occlude each other). This leads to tracking applications where a strong intra-group occlusion causes the loss of that group.

Co-PF solves this problem, and permits a very fine estimation of the whereabouts of a scene, making the group tracking very robust.

---

[1]Please note that the ground plane position is inferred employing the calibration of the camera.

[2]This is analogue to what was done in [7] for the single individuals.

[3]Detection, split, and merge of groups are not considered here, because the focus of this paper is mainly on the novel collaborative mechanism.

It basically injects the information collected by the MOT into the MGT. Considering the filtering expression in Eq. 4, the fusion occurs on the posterior at time $t - 1$:

$$p(X_{t-1}^g | z_{1:t-1}) = \int p(X_{t-1}^g, \mathbf{x}_{t-1} | z_{1:t-1}) \, d\mathbf{x}_{t-1} \quad (7)$$

$$= \int p(X_{t-1}^g | \mathbf{x}_{t-1}, z_{1:t-1}) \, p(\mathbf{x}_{t-1} | z_{1:t-1}) \, d\mathbf{x}_{t-1} \quad (8)$$

The first term of Eq. 8 is the core of our approach as it revises the group posterior distribution at time $t - 1$, also considering the states of the single individuals. In this way, the second term (the posterior at time $t - 1$ of the MOT process) may be considered as a weight that mirrors the reliability of the individual states.

A convenient way to model distributions conditioned on multiple events is that of the Mixed-memory Markov Process (MMP) [9], that decomposes a structured conditioned distribution as a convex combination of pairwise conditioned distributions. This leads to:

$$p(X_{t-1}^g | \mathbf{x}_{t-1}, z_{1:t-1}) \approx \alpha_1 p(X_{t-1}^g | \mathbf{x}_{t-1}) + \alpha_2 p(X_{t-1}^g | z_{1:t-1}), \quad (9)$$

where $\alpha_1, \alpha_2 > 0$ and $\alpha_1 + \alpha_2 = 1$. Considering Eq. 9, we can rewrite Eq. 8 as:

$$p(X_{t-1}^g | z_{1:t-1}) \approx$$

$$\alpha_1 \int p(\mathbf{x}_{t-1} | z_{1:t-1}) \, p(X_{t-1}^g | \mathbf{x}_{t-1}) \, d\mathbf{x}_{t-1} + \quad (10)$$

$$\alpha_2 \, p(X_{t-1}^g | z_{1:t-1}) \underbrace{\int p(\mathbf{x}_{t-1} | z_{1:t-1}) \, d\mathbf{x}_{t-1}}_{=1}. \quad (11)$$

At this point, it is easy to realize that $p(X_{t-1}^g | z_{1:t-1})$ becomes a combination of the natural group posterior and a marginalization of the *linking* probability $p(X_{t-1}^g | \mathbf{x}_{t-1})$, that relates a group to individuals, weighted by the MOT posterior. In other words, the group posterior is revisited by injecting in a principled way the information on the single targets (the MOT posterior), conveyed selectively by $p(X_{t-1}^g | \mathbf{x}_{t-1})$. An example will demonstrate the advantage of this formulation.

The linking probability $p(X_{t-1}^g | \mathbf{x}_{t-1})$ is factorized as an MMP as follows:

$$p(X_{t-1}^g | \mathbf{x}_{t-1}) \approx \sum_{k=1}^{K} p(X_{t-1}^g | x_{t-1}^k) \, \beta^{k,g} \quad (12)$$

$$\propto \sum_{k=1}^{K} p(x_{t-1}^k | X_{t-1}^g) \, p(X_{t-1}^g) \, \beta^{k,g} \quad (13)$$

where $\beta^{k,g} > 0 \; \forall k, g$ and $\sum_k \beta^{k,g} = 1$. Each term of the sum in Eq. 12 represents the posterior probability that the gth group X_{t-1}^g contains the kth target x_{t-1}^k.

In Eq. 13, the posterior is modeled employing the Bayes rule, where $p(x_{t-1}^k | X_{t-1}^g)$ defines the *linking* likelihood that each single individual state x_{t-1}^k is a subpart of X_{t-1}^g. Hence, we define a probability model based on three components: 1) appearance similarity, 2) dynamics consistency, and 3) group membership. The appearance similarity is encoded by the Bhattacharyya distance between the HSV histograms of the two entities: $d_{\text{HSV}}(X_{t-1}^g, x_{t-1}^k)$. The dynamics consistency rewards the person state whose motion component is similar to that of the group. In practice, we check the 2D displacement on the floor by calculating $d_{\text{dir}}(X_{t-1}^g, x_{t-1}^k) = $
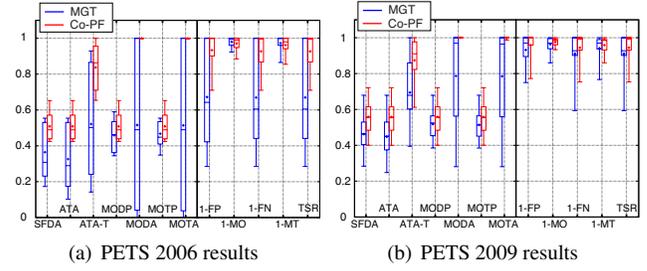

(a) PETS 2006 results   (b) PETS 2009 results

Fig. 2. Statistics on the synthetic test set.

$|1 - |\text{dir}(X_{t-1}^g) - \text{dir}(x_{t-1}^k)|/\pi|$, where $\text{dir}(\cdot)$ gives the direction (an angle) of the person or group. Finally, the group membership evaluates the spatial proximity of the person state and of the group state:

$$d_{\text{mbr}}(X_{t-1}^g, x_{t-1}^k) = \begin{cases} 1 & \text{if} \quad x_{t-1}^k \in X_{t-1}^g \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

where the membership operator $\in$ controls if the kth person position is inside the gth group ellipse. Therefore, $p(x_{t-1}^k | X_{t-1}^g) = d_{\text{HSV}}(X_{t-1}^g, x_{t-1}^k) \cdot d_{\text{dir}}(X_{t-1}^g, x_{t-1}^k) \cdot d_{\text{mbr}}(X_{t-1}^g, x_{t-1}^k)$. The coefficients $\beta^{k,g}$ express a linking preference that an object belongs to a group, and are left here as uniform, i.e., $\beta^{k,g} = 1/G$.

Finally, the prior $p(X_{t-1}^g)$ discards the biggest and the smallest group hypotheses, rejecting the particles in which the size of the group is below a threshold τ_b or above a threshold τ_a.

An example that explains the strength of our formulation can be represented by an intra-group occlusion in the gth group at time $t - 1$, which is very common due to the dynamical nature of a group of moving people. Let x_{t-1}^k a target of the group X_{t-1}^g that vanishes as occluded by the remaining individuals of that group. The group posterior $p(X_{t-1}^g | z_{1:t-1})$ will not be very high, for the limits of the visual, rigid, group representation. However, the MOT process, dealing with single objects and managing their occlusions, will "understand" the fact that x_{t-1}^k is occluded, producing a high $p(x_{t-1}^k | z_{1:t-1})$. This probability value will flow through $p(X_{t-1}^g | \mathbf{x}_{t-1})$, which is high because, even if occluded, the position and the velocity of x_{t-1}^k are correctly estimated by the MOT process, and will give a high linking likelihood. This will reinforce the final estimation of the hybrid posterior for X_{t-1}^g, thus permitting to estimate the subsequent group sample set in a more correct way.

4. EXPERIMENTAL RESULTS

Our approach has been evaluated on synthetic data and publicly available datasets (PETS 2006[4] and PETS 2009[5]). We carried out a comparative analysis wrt the MGT (without the proposed collaboration stage), highlighting that Co-PF is more able to deal with intra- and inter-group occlusion. Other approaches have not been taken into account because of the lack of: 1) on-line available code for any of the approaches in the state of the art 2) a shared, labelled, dataset.

The simulations on the synthetic test set[6] are carried out, in order to build statistics on ground-truthed sequences. The test set is built to emulate the scenarios in PETS dataset by using the same background and the same calibration data. Each sequence contains static images of people walking in the environment and forming groups.
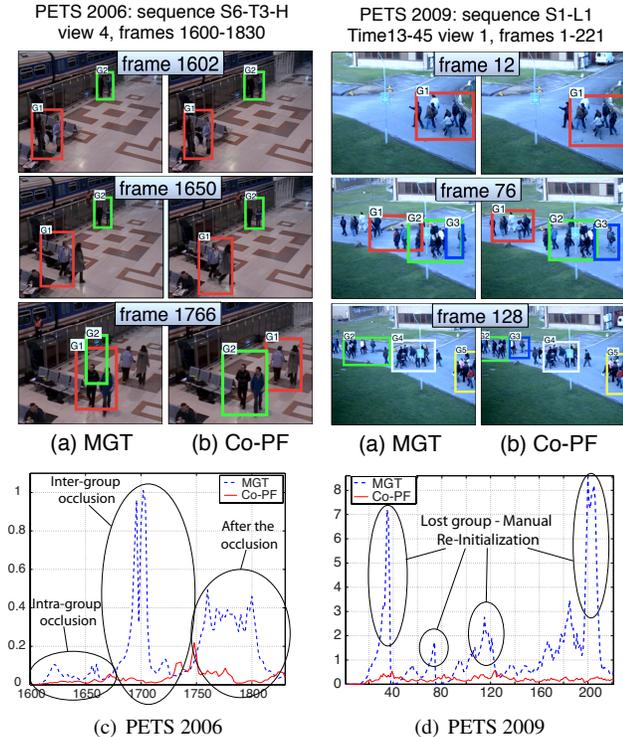
Fig. 3. Comparison of MGT (first and third column) and Co-PF (second and fourth column) on PETS 2006 and PETS 2009. The second row compares the PF uncertainty [12] in the two experiments.

We artificially create a set of 26 sequences (13 for each dataset), choosing two different points of view in order to deal with variably scaled people: the first camera is closed to the people, while the second one is far. The number of people and the number of groups vary in different sequences from 3 to 20 and from 1 to 5, respectively. The number of person in a group varies from 2 to 6. The parameters are set as follows: $\sigma_\mu = 0.05$, $\sigma_\lambda = 0.05$, $\sigma_\theta = \pi/40$, 256 bin are used for the HSV histogram, $\alpha_1 = \alpha_2 = 0.5$, $\tau_b = 0.5$, $\tau_a = 2.5$.

A comparison has been done between the Co-PF with $N = 50$ and $N_g = 50$ (the number of particles for each group) and MGT with $N'_g \approx N_g + N \cdot \frac{K^2}{G^2 \cdot C}$, where $C = 5$ has been empirically chosen, K and G are the number of people and groups, respectively. In this way, the computational burden of the two methods is similar. To evaluate the performance on the synthetic test set, we adopt the follow measures: Average Tracking Accuracy (ATA), Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), False Positive (FP), Multiple Objects (MO), False Negative (FN), Tracking Success Rate (TSR) and so on (further details in [10, 11]). For each measure, a *boxplot* representation is given [10], where the box is defined by the 1st quartile, median, and the 3rd quartile; the extremities outside the box are the smallest and largest value, and the "+" is the mean value. The comparison (Fig. 2(a) and Fig. 2(b)) shows that in the PETS2006 synthetic dataset our Co-PF strongly outperforms the MGT in terms of all the measures. Even though the PETS2009 sequences are slightly harder, Co-PF often succeeds where MGT fails, yielding to higher performances.

Moreover, we perform the test on portions of the PETS datasets, using the same settings. We consider sequences where the groups were not subjected to splits or merges, in order to stress the capability of tracking group entities with intra- and inter-group occlusions. Initialization of groups has been done by fitting the μ^g and Σ^g to the projections of the individuals new entries on the ground plane.

If lost, a group is manually reinitialized. We show here two representative examples. In real scenarios, MGT is not able to deal completely with the intra- and inter-group dynamics (Fig. 3(a)). On the other hand, Co-PF exploits the MOT results, enriching the posterior knowledge given by the MGT (Fig. 3(b)).

To give further support to our Co-PF, we evaluate the uncertainty of the particle filters [12]. Fig. 3(c) depicts that the MGT uncertainty is peaked when an intra- and inter-group occlusion occurs. After the occlusion the uncertainty is high because the track is erroneously lost (two tracks on a single group). Fig. 3(d) shows a similar behavior of Fig. 3(c), highlighting that the MGT looses the tracks several times.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose and extensively evaluate a collaborative framework for group tracking. Two processes are involved: the group process simplifies groups as atomic entities, dealing successfully with occluding, multiple groups. The individual process captures how individuals move, refining the estimations done by the group process directly in the posterior distribution. In this way, the group tracking process can evolve in a more robust way. Future directions will focus on dealing with split and merge dynamics, which are ignored here, but that can be easily embedded in our framework, in a seamless way.

6. REFERENCES

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

[2] J. S. Marques, P. M. Jorge, A. J. Abrantes, and J. M. Lemos, "Tracking groups of pedestrians in video sequences," in *Workshop at CVPR*, 2003, vol. 9, pp. 101–101.

[3] G. Gennari and G.D. Hager, "Probabilistic data association methods in visual tracking of groups," in *CVPR*, 2004, vol. 2, pp. 876–881.

[4] T. Mauthner, M. Donoser, and H. Bischof, "Robust tracking of spatial related components," in *ICPR*, 2008, pp. 1–4.

[5] Wen-Chieh Lin and Yanxi Liu, "A lattice-based mrf model for dynamic near-regular texture tracking," *IEEE TPAMI*, 29(5):777–792, 2007.

[6] Y. Lao and F. Zheng, "Tracking a group of highly correlated targets," in *ICIP*, 2009.

[7] O. Lanz, "Approximate bayesian multibody tracking," *IEEE TPAMI*, 28(9):1436–1449, 2006.

[8] Edward T. Hall, *The Hidden Dimension*, Anchor Books, 1966.

[9] L. K. Saul and M. I. Jordan, "Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones," *Mach. Learn.*, 37(1):75–87, 1999.

[10] R Kasturi, D Goldgof, P Soundararajan, V Manohar, J Garofolo, R Bowers, M Boonstra, V Korzhova, and J Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE TPAMI*, 31(2):319–336, 2009.

[11] K. Smith, D. Gatica-Perez, J. Odobez, and S. Ba, "Evaluating multi-object tracking," in *CVPR*, 2005, pp. 36–43.

[12] E. Maggio, F. Smerladi, and A. Cavallaro, "Combining colour and orientation for adaptive particle filter-based tracking," in *BMVC*, 2005.