Improving multiple pedestrians tracking with semantic information

Jorge Francisco Madrigal Diaz, Jean-Bernard Hayet, Frédéric Lerasle

Signal, Image and Video Processing manuscript No.
(will be inserted by the editor)

Improving multiple pedestrians tracking with semantic information

Francisco Madrigal · Jean-Bernard Hayet · Frédéric Lerasle

Received: date / Accepted: date

Abstract This work presents an interacting multiple pedestrian tracking method for monocular systems that incorporates a prior knowledge about the environment and about interactions between targets. Pedestrian motion being ruled by both environment and social aspects, we model these complex behaviors by considering 4 cases of motion: going straight; finding one's way; walking around and standing still. They are combined within an Interacting Multiple Model Particle Filter strategy. We model targets interactions with social forces, included as potential functions in the weighting process of the Particle Filter. We use different social force setups within each motion model to handle high level behaviors (collision avoidance, flocking...). We evaluate our algorithm on challenging datasets and show that such semantic information improves the tracker performance compared to the literature.

Keywords Pedestrian visual tracking · Particle filter · Social forces · Semantic based tracking

1 Introduction

Multi-object tracking (MOT) has been a very active research area in recent years. The techniques developed for MOT have found applications in the automatization of processes in areas such as robotics or video surveillance, among others. The main two ingredients of most MOT techniques are (1) the modeling of the target visual appearance, and (2) the modeling of the prior knowledge about the targets motion. In this work, we focus on the latter, i.e., the use of probabilistic models for explaining the observed motions and the interactions between pedestrians in a scene captured by a video-surveillance camera. Although, at first sight, the nature of pedestrian motion may look quite chaotic, studies [12, 19] have shown that pedestrian behavior is strongly influenced by the context, namely the other pedestrians in his surroundings and, beyond, the environment configuration and its clutter. This has been the starting point for cornerstone research in the modeling of group behaviors, i.e., for the design of escape routes in public spaces. As an example, consider the persons present in Fig. 1. The couple at the center of the image is standing in place, while other pedestrians are moving around, in groups or alone, with different velocities. All agents velocities are clearly influenced by other agents intentions and proximity: They may want to avoid the obstacle made by the couple, or enter into conversation. To model this behavioral context, global positions, orientations, and velocities of the targets are natural variables to be used. For instance, pedestrians in the same group should have similar orientations, whereas two nearby people talking to each other should be oriented in an opposite way. This kind of semantic interactions are not used in most tracking approaches. However, our claim is that the inference of the pedestrians interactions based on semantic dynamic models could improve the tracking performance by producing better predictions in stochastic filtering. In this paper, we consider a simplified model of four cases of motions (one probabilistic model per motion), obtained by the analysis of the pedestrians in a mall [19]. For each of these motion models, we include the modeling of target interactions through potential functions, encoding the concept of social forces. Finally, our motion models are integrated in one single framework with

F. Madrigal · J-B Hayet
Centro de Investigación en Matemáticas (CIMAT),
Guanajuato, Gto., México
Tel.: +52 473 732 7155
E-mail: {pacomd,jbhayet}@cimat.mx

F. Lerasle
Univ. de Toulouse, UPS, CNRS-LAAS, 7 avenue colonel Roche,
F-31400 Toulouse, France.
E-mail: lerasle@laas.fr

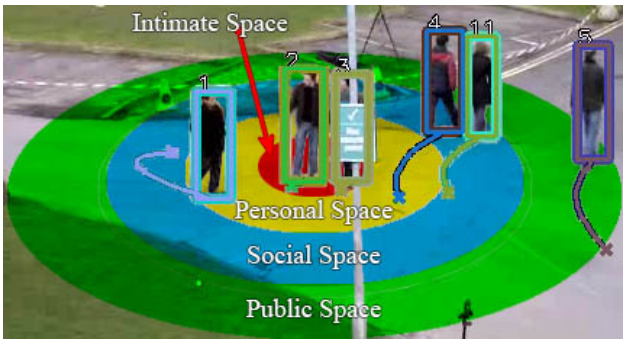

Fig. 1 Pedestrians with multiple motion dynamics. The interaction of the person in the middle of the image with others depends on the region that they occupy. From proxemic theory, these regions can be divided in four: Intimate (Red), Personal (Yellow), Social (Blue) and Public (Green) space.

the Interacting Multiple Model scheme under a Particle Filter methodology [6], coined as IMM-PF.

In the work presented here, the motion models are developed with semantic information modeled as in [19], that allows to handle in a more natural way the human walking in sparsely crowded scenes.

We propose a decentralized tracking framework, i.e., one filter dedicated to each target. Even if the trackers are essentially individual, they share semantic information through a prior knowledge about the expected social behavior in each motion model among a set of competing models. Our modeling considers the body pose of each target (in the same vein as [8]) as a feature to control these interactions. We demonstrate that our proposal outperforms existing approaches thanks to large scale comparative evaluations. The Fig. 2 provides an overview of our proposal.

The structure of the paper is as follows: Section 2 discusses related work. The general formulation of our proposed IMM-PF is presented in Section 3. The Section 4 describes our contribution in the modeling of the pedestrian behavior (motion and interaction). Results are presented in Section 5. Finally, conclusions are drawn in Section 6.

2 Related work

Most of the time, naive dynamic models are used as priors in MOT frameworks, i.e., the constant velocity model [8,5], random walks [15], target detector output [5], among others.

Unfortunately, those models are rough approximations of the real dynamic of the targets and they lack semantic information that could improve tracking performance by identifying common group walking patterns, for example. [15] proposes a technique to model a simple kind of interaction between individual trackers. They use a potential function to give more weight to those particles of a particle filter that are far from other trackers, helping to keep the trackers apart. However, this method can not be extended very well to multiple behaviors since the interaction models may contradict each other. In [5], the authors present a framework to track individuals and groups of pedestrians at the same time, using semantic information about the group formation. However, no motion prior information is used. On the other hand, [10] makes use of semantic information to identify groups from independent trackers. [18] introduces a multi-camera tracking system with non-overlapping field of view. It uses a social force model to generate multiple hypothesis about the movements of a non-observed target who has left the field of view of a camera. Those hypothesis are considered for target re-identification. [23] solves the tracklet data association problem as a directed graph, by weighting the edges according to some social conditions. In [20], the targets interact in such a way that they choose a free collision trajectory. To this end, this work finds the optimal next position of all trackers based on an energy function that considers the targets future position, desired speed and final destination. Other MOT systems consider trackers interactions only during the detector association stage [7], or only when targets touch each other, or when one is occluded in one camera. The objective in that case is to avoid the coalescence phenomenon and to solve the data association problem.

Capturing the complex behavior of targets like pedestrians can be really challenging. An elegant solution is to rely on a mixture of motion models through the Interacting Multiple Model (IMM) methodology. IMM maintains a pool of distinct, competing models and weights each of them according to its importance in the posterior distribution [6, 13]. In [13], target tracking is simulated with a bank of Kalman Filters, where each filter is associated to a distinct linear motion model, within the IMM methodology. This proposal is fast and suitable for a large number of targets. In [22], a similar bank of filters was employed in a hybrid foreground subtraction and pedestrian tracking algorithm. It uses the tracking result as a feedback to improve the foreground subtraction. [14] proposes another Kalman-based IMM for pedestrian tracking which is similar to ours, with two classic motion models: constant position and constant velocity, to track a few targets.

However, the Kalman filter cannot use non-linear models and the IMM schemes based on it can not recover when one filter of the bank fails. [6] proposes an IMM implementation with Particle Filter (that we will refer to as IMM-PF). They associate a fixed number of particles to each model and weight the models according to their importance in the filter. This proposal suffers from a waste of computational resources when processing many particles with low importance models. In [16], each particle motion model has the possibility of evolving over time, passing from a *moving* to a *stopped* state. Those changes are handled with a transition

matrix of fixed probability values. However, those fixed values can not represent faithfully how the real model changes.

Contributions. To overcome the limitations of the common naive dynamic models (widely used in MOT [21,14,15]), we propose a decentralized tracking system with a motion model that considers semantic information to improve pedestrian tracking. We model this high level pedestrian behavior at two levels: motion and interaction. We emulate the complex pedestrian motion with Interactive Multiple Models (IMM), developed from observation analysis [19]. We expand the work of Khan et al. [15] to multiple pedestrian tracking and include more realistic interaction between trackers coming from the simulation community, known as social forces. We demonstrate, in several challenging video sequences through both qualitative and quantitative evaluations, that such semantic information improves the tracking performance compared to conventional approaches in the literature.

3 Particle Filter-Interacting Multiple Models

We formulate the tracking problem in a Bayesian framework, where we infer the target state $\mathbf{X}$ at time t ($\mathbf{X}_t$) given the set of observations $\mathbf{Z}_{1:t} \stackrel{\text{def}}{=} \{\mathbf{Z}_1 \dots \mathbf{Z}_t\}$. Under the Markov assumption, the posterior is estimated recursively:

$$\begin{cases} p(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int p(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) d\mathbf{X}_{t-1}, \\ p(\mathbf{X}_t|\mathbf{Z}_{1:t}) \propto p(\mathbf{Z}_t|\mathbf{X}_t) p(\mathbf{X}_t|\mathbf{Z}_{1:t-1}). \end{cases} \quad (1)$$

The Bayes filter of Eq. 1 includes prediction (first row) and correction (second row) steps. Following the IMM strategy [6], our motion model $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is a mixture of M distributions as:

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = \sum_{m=1}^{M} \pi_t^m p^m(\mathbf{X}_t|\mathbf{X}_{t-1}), \quad (2)$$

where the terms π_t^m weigh each model contribution in the mixture. Thus, the posterior of Eq. 1 is reformulated as:

$$\begin{cases} p(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int \sum_{m=1}^{M} \pi_t^m p^m(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) d\mathbf{X}_{t-1}, \\ p(\mathbf{X}_t|\mathbf{Z}_{1:t}) \propto p(\mathbf{Z}_t|\mathbf{X}_t) p(\mathbf{X}_t|\mathbf{Z}_{1:t-1}). \end{cases} \quad (3)$$

Since the contribution weight does not depend on the previous state $\mathbf{X}_{t-1}$, we move this term out of the mixture distribution. Hence, the filter of Eq. 3 is rewritten as:

$$p(\mathbf{X}_t|\mathbf{Z}_{1:t}) \propto \sum_{m=1}^{M} \pi_t^m p(\mathbf{Z}_t|\mathbf{X}_t) p^m(\mathbf{X}_t|\mathbf{Z}_{1:t-1}), \quad (4)$$

with $p^m(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int p^m(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) d\mathbf{X}_{t-1}$. The terms π_t^m are updated in function of their respective likelihoods [6]: $\pi_t^m = \pi_{t-1}^m \int p(\mathbf{Z}_t|\mathbf{X}_t) p^m(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) d\mathbf{X}_t$. The particle filter approximates the posterior in Eq. 4 by a set of N weighted samples or particles. The multi-modality is implemented by assigning one motion model to each particle, indicated by a label $l \in \{1 \dots M\}$. Thereby, a particle n at time t is represented by $(\mathbf{X}_t^{(n)}, \omega_t^{(n)}, l^{(n)})$.

In the IMM-PF methodology, the model $m = \{1 \dots M\}$ contributes to the posterior estimation according to its importance, which is defined by a weight π_t^m. Each model m has N_m particles associated to it, with a total of $N = \sum_{m=1}^{M} N_m$ particles. The posterior is represented by considering both particles weights ($\omega_t^{(n)}$) and models weights (π_t^m):

$$\begin{aligned} &p(\mathbf{X}_t|\mathbf{Z}_{1:t}) = \sum_{m=1}^{M} \pi_t^m \sum_{n \in \psi_m} \omega_t^{(n)} \delta_{\mathbf{X}_t^{(n)}}(\mathbf{X}_t), \\ &\text{s.t. } \sum_{m=1}^{M} \pi_t^m = 1 \text{ and } \sum_{n \in \psi_m} \omega_t^{(n)} = 1, \end{aligned} \quad (5)$$

where $\psi_m \stackrel{\text{def}}{=} \{n \in \{1 \dots N\} : l^{(n)} = m\}$ represents the indices of the particles that belong to model m.

3.1 Sampling and dynamic model

We use an importance proposal distribution $q(\cdot)$, that approximates $p(\mathbf{X}_t|\mathbf{X}_{t-1}, \mathbf{Z}_{1:t})$, and from which we can draw samples. In the multiple motion model case, we have M proposals, such as: $\mathbf{X}_t^m \sim q^m(\mathbf{X}_t|\mathbf{X}_{t-1}, \mathbf{Z}_{1:t})$. Here, we sample a new state for each particle from the motion model corresponding to its label $l^{(n)}$. This model is supposed to be a Gaussian distribution $\mathcal{N}(\mathbf{X}_t; tr_{l^{(n)}}(\mathbf{X}_{t-1}^{(n)}), \Sigma_{l^{(n)}})$, where $tr_{l^{(n)}}(\cdot)$ is the deterministic form of the motion model (which will be detailed in the next section). The index $l^{(n)}$ indicates the model the particle n follows.

3.2 Observation model and correction step

We implement a probabilistic observation model $p(\mathbf{Z}_t|\mathbf{X}_t)$ inspired from [21,8]. [21] relies on HSV-space color and motion histograms. We define a reference histogram h_{ref} anytime we create a new tracker. The likelihood is evaluated between h_{ref} and the current histogram $h^{(n)}$ (corresponding to $\mathbf{X}_t^{(n)}$) through the Bhattacharya distance. We include spatial information with the color observation by using multiple-region reference models (two histograms per target, one for the top part of the person and another for the bottom part) as it has been shown to be more robust [21].

Following [8], we also include observations related to the target orientation, because, as it will be explained, orientation is part of our state, as an angle θ_t. It is discretized into eight directions. The body pose angle is evaluated with a set of multiple-level Histogram of Oriented Gradients features (HOG) $f^{(n)}$ extracted from the image inside each $\mathbf{X}_t^{(n)}$. They are decomposed into a linear combination of O training samples $\mathbf{F} = \{f_1, \dots, f_O\}$: $f^{(n)} \approx a_1 f_1 + \dots + a_O f_O = \mathbf{Fa}$, where $\mathbf{a} = \{a_1, \dots, a_O\}$ is the weights vector subject to non-negative constraints ($a_o \geq 0$ for $o \in [1, O]$). Each sample has associated a label $l'_o \in \{1 \dots 8\}$ corresponding to its orientation. The idea is to find an optimal decomposition of the detected features in terms of the training samples, i.e., to determine a set of positive weights ($\mathbf{a}^*$) such that:

$$\mathbf{a}^* = \arg\min \|f^{(n)} - \mathbf{Fa}\|_2^2 + \lambda \|\mathbf{a}\|_1,$$

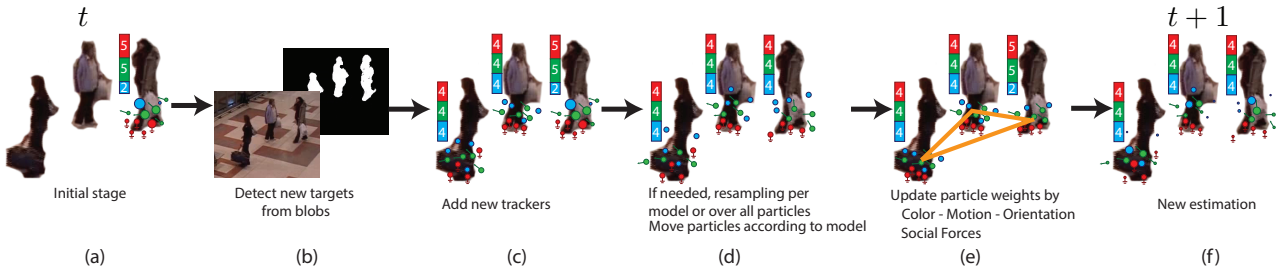

Fig. 2 Workflow of our proposal. (a) Initial stage: the tracker system at a given time t. (b) Input image is used to detect pedestrian blob candidates. (c) A new tracker is created for each isolated blob candidate. The circles represent the particles, and their color and diameter depict the motion model id and weight, respectively. The left bar shows the number of particles that each model has. (d) IMM-PF prediction step: a resampling (per model or over all particles) is applied if needed; the particles are moved according to its model. (e) IMM-PF correction step: particles weights are updated from color, motion and orientation cues; the social force model is applied to each interacting trackers. (f) Final tracker estimation.

where λ controls the regularization. Then, the orientation likelihood $p_\theta(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$ is calculated as the normalized sum of the weights of $\mathbf{a}^*$:

$$p_\theta(\mathbf{Z}_t|\mathbf{X}_t^{(n)}) = \frac{1}{\|a^*\|_1} \sum_{o \in \rho_t(\theta_t^{(n)})} a_o^*,$$

where $\rho_t(\theta_t^{(n)})$ is the set of indexes o of the images from the training database whose labels l'_o have the same (discretized) orientation $\theta_t^{(n)}$ as the particle n. Assuming independence between the observation components (color cue, motion cue, orientation cue), the likelihood of the observation $\mathbf{Z}_t$ evaluated at the state of particle n is defined as the combination of the three models:

$$p(\mathbf{Z}_t|\mathbf{X}_t^{(n)}) = p_c(\mathbf{Z}_t|\mathbf{X}_t^{(n)}) p_m(\mathbf{Z}_t|\mathbf{X}_t^{(n)}) p_\theta(\mathbf{Z}_t|\mathbf{X}_t^{(n)}),$$

where $p_c(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$ and $p_m(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$ are the color and motion cues[21], respectively, and $p_\theta(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$ is the orientation likelihood described above. Thus, particles weights are updated by:

$$\omega_t^{(n)} = \frac{\tilde{\omega}_t^{(n)}}{\sum_{i \in \psi_m} \tilde{\omega}_t^{(i)}}, \quad \tilde{\omega}_t^{(n)} = \frac{\omega_{t-1}^{(n)} p(\mathbf{Z}_t|\mathbf{X}_t^{(n)}) p^{l^{(n)}}(\mathbf{X}_t^{(n)}|\mathbf{X}_{t-1}^{(n)})}{q^{l^{(n)}}(\mathbf{X}_t^{(n)}|\mathbf{X}_{t-1}^{(n)}, \mathbf{Z}_{1:t})}. \quad (6)$$

By assuming that the proposal and prior distribution are the same, we have:

$$\tilde{\omega}_t^{(n)} = \omega_{t-1}^{(n)} \cdot p(\mathbf{Z}_t|\mathbf{X}_t^{(n)}), \quad (7)$$

$$\pi_t^m = \frac{\pi_{t-1}^m \tilde{\omega}_t^m}{\sum_{i=1}^M \pi_{t-1}^i \tilde{\omega}_t^i}, \quad \tilde{\omega}_t^m = \sum_{j \in \psi_m} \tilde{\omega}_t^{(j)}. \quad (8)$$

Thus, Eqs. 6 and 8 ensure that the constraints on Eq. 5 are always satisfied.

3.3 Resampling

We implement the resampling process as in [17] (Fig. 2-d). It performs the sampling, if needed, in one of two ways:

1. A sampling over all particles, following a common Cumulative Distribution Function built with the weights of particles $\omega_t^{(n)}$ and models π_t^m. The best particles from the best models are sampled more often, leaving more particles with models fitting better the target motion.
2. A sampling on a per model basis. Each model keeps a minimum of $\gamma \stackrel{\text{def}}{=} 0.1 * N$ particles to preserve diversity. If the model has less particles than a threshold ($N_m < \gamma$), we draw new samples from a Gaussian distribution: $\mathcal{N}(\bar{\mathbf{X}}_{t-1}, \mathbf{S}_{t-1})$, where $\bar{\mathbf{X}}_{t-1}$ and $\mathbf{S}_{t-1}$ are the weighted mean and covariance of all particles of the previous distribution. We take less samples from the models with more particles, to leave the total number of particles N unchanged. This resampling manages the model transition implicitly, so no prior transition information is required.

The resampling over all particles is applied every 4 frames and the one over models is applied every 5 frames.

4 Models for pedestrian semantic behavior

This section describes our main contribution with more details. We propose a multiple-motion model that improves the tracking performance by fitting better to different pedestrian dynamics (Fig. 2-e). Also, it incorporates semantic information about the interaction of the targets, with a set of expected behavioral rules relying on the concept of interpersonal space between targets (illustrated by Fig. 1).

The target state is defined as a bounding box, including its position in the image plane (x, y), its global shoulders orientation θ, and its linear and angular velocities (v_l, v_θ). Hence, the state $\mathbf{X}$ stands as $(x, y, \theta, v_l, v_\theta)^T$. The bounding box dimensions (h, w) around the pedestrians are fixed according to the average size of an adult person, given the camera projection matrix, at the specified image location (see [17]). As we have already mentioned it, the reason why we also include the orientation is that target interactions are common in MOT, and that the orientation is strongly correlated to the pedestrian's "intentionality" (characterized by the shoulders orientation), i.e., pedestrians from the same group share similar orientations.

4.1 Priors on pedestrian dynamics

According to [19], four pedestrian motions can be considered in human-centered environment:

- **Going straight.** The pedestrians go directly to their goal, as fast as possible, with small variations in the trajectory.
- **Finding one's way.** The pedestrians have an approximate idea of their destination (i.e., an address over a route). They walk at a regular speed, with more variations in their trajectories.
- **Walking around.** The pedestrians don't have a specific goal. They walk at slow speed and tend to change their trajectories more often.
- **Stand still.** The pedestrians remain at the same position, occasionally changing their body orientation. They may be interacting with other persons.

We build 4 motion models to emulate those behaviors. The first three cases ($k = 1, 2, 3$) are associated to the following generic transition model:

$$tr_k(\mathbf{X}) = \begin{bmatrix} x + v_l * \cos(\theta) \\ y + v_l * \sin(\theta) \\ \theta + v_\theta \\ \mu_k \\ v_\theta \end{bmatrix} + \begin{bmatrix} \mathcal{N}(0, \sigma_x) \\ \mathcal{N}(0, \sigma_y) \\ \mathcal{N}(0, \alpha(v_l) * \sigma_\theta) \\ \mathcal{N}(0, \sigma_{v_l,k}) \\ \mathcal{N}(0, \alpha(v_l) * \sigma_{v_\theta,k}) \end{bmatrix}, \quad (9)$$

where σ_x, σ_y and σ_θ are predefined constant values and represent a variance of $0.2m$, $0.2m$ and 5 degrees respectively. The new position is updated as a constant velocity non-holonomic motion model. Normally, a pedestrian who walks fast has a rather constant orientation. Following this idea, we calculate the new orientation and angular velocity by considering an adapting level of noise, controlled by $\alpha(v) = \exp(-v^2/\sigma_\alpha)$. Hence, the higher the linear velocity v_l, the smaller the variance of the Gaussian noise. The μ_k and $\sigma_{\cdot,k}$ values depend on the model to be used, allowing to control the behavior of the aforementioned categories 1, 2 and 3. These parameters are estimated following the algorithm of section 4.2. The **stand still** case is simpler:

$$tr_4(\mathbf{X}_t) = \begin{bmatrix} I_{3\times3} & 0_{3\times2} \\ 0_{2\times3} & 0_{2\times2} \end{bmatrix} \mathbf{X}_t + \nu, \quad (10)$$

where ν is a realization of a Gaussian noise. Pedestrians are also influenced by a set of external rules known as social forces (SF) [12]. Those SF depend on the dynamics of the people. They will be detailed in Section 4.3.

4.2 Tuning of the free parameters

In section 4.1, we described a transition model (Eq. 9) that incorporates semantic information about the pedestrian motions. This model is controlled by three parameters: the mean μ_k and the variance $\sigma_{v_l,k}$ of the target speed, and the variance in the pedestrian orientation $\sigma_{v_\theta,k}$. For the three presented cases, we estimate those parameters as follows. Initially, we set them with the values proposed in [19] for pedestrians in a shopping mall. To make our framework more adaptable to other scenarios, we estimate those parameters by using the Particle Marginal Metropolis-Hastings (PMMH) algorithm [4]. This algorithm is a Markov Chain Monte Carlo (MCMC) algorithm that recovers jointly the state $\mathbf{X}_t$ and the model parameters $\beta \stackrel{\text{def}}{=} \{\mu_k, \sigma_{v_l,k}, \sigma_{v_\theta,k}\}$. In a Bayesian context, the parameters follow a prior distribution $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$, where μ_β is set according to the parameter values presented in [19] and $\sigma_\beta = 0.5$. The idea is to estimate their posterior $p(\beta|\mathbf{Z}_{1:t})$ following the Metropolis-Hastings strategy. At an iteration g, a candidate β^c is generated from a proposal distribution $q_\beta(\beta^c|\beta_{g-1}) \sim N(\beta^c; \beta_{g-1}, 0.5)$. Then, we apply the filter from section 3 with the parameters β^c. This candidate is accepted with probability:

$$\min\left\{1, \frac{\hat{p}(\mathbf{Z}_{1:t}|\beta^c)\kappa(\beta^c)}{\hat{p}(\mathbf{Z}_{1:t}|\beta_{g-1})\kappa(\beta_{g-1})} \frac{q_\beta(\beta_{g-1}|\beta^c)}{q_\beta(\beta^c|\beta_{g-1})}\right\}$$

where $\hat{p}(\mathbf{Z}_{1:t}|\beta^c) = \frac{1}{N}\sum_n^N \tilde{\omega}_t^{(n)}$ is the particle filter unbiased estimate of the marginal likelihood. Note that this quantity is estimated with the particle weights of Eq. 7.

4.3 Social behaviors for trackers interaction

The social forces (SF) model makes possible to model the interaction between trackers. We associate a set of SFs to each motion model according to the behavior expected in each case. These behaviors are selected from the proxemic theory [11] and depends on the space occupied by the interacting trackers. In Fig. 1, we depict an example, where the central pedestrian (labeled as 2) interacts with his neighbors according to their relative position (circles of colors). The state $\mathbf{X}_t$ is projected into the world plane to control the effect of each force in real coordinates. We use two SFs: (1) A repulsion force, keeping the trackers apart from each other, and preventing identity switching or collisions; (2) An attraction force, keeping the targets close to each other, and modeling social groups. By setting both forces with different values, we can model many kinds of behaviors.

Interactions are modeled with pairwise potential functions [15]. We define one such potential, for each of the M models, $SF_m(\mathbf{X}_i, \mathbf{X}_j)$ which can be easily included in the prior motion model of Eq. 2:

$$p(\mathbf{X}_{t,i}|\mathbf{X}_{t-1,i}) = \sum_{m=1}^{M} \pi_t^m p^m(\mathbf{X}_{t,i}|\mathbf{X}_{t-1,i}) \prod_{j\in\varphi_i} SF_m(\mathbf{X}_{t,i}, \mathbf{X}_{t,j}),$$

where $\varphi_i \stackrel{\text{def}}{=} \{j \in \{1 \dots N\} : i \neq j\}$. As in Eq. 3, the interaction term $SF_m(\cdot)$ does not depend on the previous state $\mathbf{X}_{t-1}$, so, this term is moved out of the mixture distribution

with π_t^m. This way, the posterior of Eq. 4 for a target i is reformulated as:

$$p(\mathbf{X}_{t,i}|\mathbf{Z}_{1:t}) \propto \sum_{m=1}^{M} \pi_t^m p(\mathbf{Z}_t|\mathbf{X}_{t,i}) \cdot \prod_{j\in\varphi_i} SF_m(\mathbf{X}_{t,i},\mathbf{X}_{t,j}) p^m(\mathbf{X}_{t,i}|\mathbf{Z}_{1:t-1}).$$

Since the interaction term is out of the mixture distribution, we can treat it as an additional factor in the importance weight. Thus, we weight the samples of Eq. 7 according to:

$$\tilde{\omega}_{t,i}^{(n)} = \omega_{t-1,i}^{(n)} \cdot p(\mathbf{Z}_t|\mathbf{X}_{t,i}^{(n)}) \prod_{j\in\varphi_i} SF_{l_i^{(n)}}(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}),$$

where $\hat{\mathbf{x}}_t = \left[\hat{x}, \hat{y}, \hat{\theta}, \hat{v}_l, \hat{v}_\theta\right]_t^T$ is the state projected on the ground plane through the homography (which let us measure the targets real positions), and $\hat{r} = [\hat{x}, \hat{y}]^T$ is the position. The term $SF_{l_i^{(n)}}(\cdot,\cdot)$ is the social force model the particle n is associated to. We measure the distance between two trackers (i, j) through the L2 norm as $\hat{d}_{i,j} = \|\hat{r}_{i,t} - \hat{r}_{j,t}\|$. All the distance considerations in the rest of the paper come from the study of nonverbal communication known as proxemics and try to emulate the notion of personal space depicted in Fig. 1. We define the social forces for each motion models as:

1. **Going straight.** The pedestrians who walk fast are aware of the obstacles present in their public space (green circle in Fig. 1) and decide with enough anticipation their direction for a comfortable free-collision path. In that case, we use a repulsion function over any tracker under a public distance, i.e., $\hat{d}_{ij} < PD$, depicted as green circles in Fig. 1. The social force for case 1 (sec. 4.1) is:

$$SF_1(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,\varphi_i}) = \prod_{j\in\varphi_i} GS(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}), \tag{11}$$

$$GS(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} 1 - \exp\left(-\dfrac{d_{i,j}^2}{\sigma_{f_1}^2}\right) & \text{if } \hat{d}_{i,j} < 3.5m, \\ 1 & \text{otherwise.} \end{cases}$$

We have used $PD = 3.5m$ and $\sigma_{f_1} = 2m$.

2. **Finding one's way.** The pedestrian walks at middle/high speed, moving alone, inside a group or merges/splits from a group. At this speed, groups are not too close, preserving a social distance SD. We consider that two targets with $\hat{d}_{i,j} < SD$, $\|\hat{v}_{l,i} - \hat{v}_{l,j}\| < \epsilon_v$, and orientation $\|\hat{\theta}_i - \hat{\theta}_j\| < \epsilon_\theta$ belong to a same group. They are depicted as blue circles in Fig. 1. We model this as:

$$FW_{\texttt{attr}}(\mathbf{X}_i, \mathbf{X}_j) = \exp\left(-\frac{(\hat{d}_{i,j} - SD)^2}{\sigma_{f_2}^2}\right), \tag{12}$$

where $SD = 2.5m$ and $\sigma_{f_2} = 20cm$ is the standard deviation on distances. Otherwise, the target i evades targets j and this is modeled by:

$$FW_{\texttt{rep}}(\mathbf{X}_i, \mathbf{X}_j) = 1 - \exp\left(-\frac{d_{i,j}^2}{\sigma_{f_3}^2}\right), \tag{13}$$

with $\sigma_{f_3} = 1m$. Thus, the social force for case 2 is:

$$SF_2(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,\varphi_i}) = \prod_{j\in\varphi_i} FW(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}), \tag{14}$$

$$FW(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} FW_{\texttt{attr}}(\mathbf{X}_i, \mathbf{X}_j) & \text{if } \hat{d}_{i,j} < PD \text{ and} \\ & \|\hat{v}_{l,i} - \hat{v}_{l,j}\| < \epsilon_v \text{ and} \\ & \|\hat{\theta}_i - \hat{\theta}_j\| < \epsilon_\theta, \\ FW_{\texttt{rep}}(\mathbf{X}_i, \mathbf{X}_j) & \text{if } \hat{d}_{i,j} < PD, \\ 1 & \text{otherwise.} \end{cases}$$

3. **Walking around.** Pedestrians tend to walk at comfortable speeds, in groups. Targets belong to the same group if they satisfy $\hat{d}_{i,j} < SD$, depicted as the yellow region in Fig. 1, keeping a personal distance of QD, a similar velocity $\|\hat{v}_{l,i} - \hat{v}_{l,j}\| < \epsilon_v$ and almost the same orientation $\|\hat{\theta}_i - \hat{\theta}_j\| < \epsilon_\theta$. This flock behavior is modeled as:

$$WA_{\texttt{attr}}(\mathbf{X}_i, \mathbf{X}_j) = \exp\left(-\frac{(\hat{d}_{i,j} - QD)^2}{\sigma_{f_2}^2}\right), \tag{15}$$

where $QD = 1.5m$. Otherwise, it avoids the obstacles:

$$WA_{\texttt{rep}}(\mathbf{X}_i, \mathbf{X}_j) = 1 - \exp\left(-\frac{d_{i,j}^2}{\sigma_{f_4}^2}\right), \tag{16}$$

with $\sigma_{f_4} = 1m$. The SF influence over a particle is:

$$SF_3(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,\varphi_i}) = \prod_{j\in\varphi_i} WA(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}), \tag{17}$$

$$WA(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} WA_{\texttt{attr}}(X_i, X_j) & \text{if } \hat{d}_{i,j} < SD \text{ and} \\ & \|\hat{v}_{l,i} - \hat{v}_{l,j}\| < \epsilon_v \text{ and} \\ & \|\hat{\theta}_i - \hat{\theta}_j\| < \epsilon_\theta, \\ WA_{\texttt{rep}}(X_i, X_j) & \text{if } \hat{d}_{i,j} < SD, \\ 1 & \text{otherwise.} \end{cases}$$

4. **Stand still.** The person remains in the same position, maybe interacting with other people, i.e., talking, with an interpersonal distance of $ID = 1m$,. This is the case in Fig. 1, where the target 2 speaks with target 3. We model this behavior with an attraction function between two close trackers ($\hat{d}_{i,j} < QD$) with opposite orientations ($\hat{\theta}_{i,j} = \|\hat{\theta}_i - \hat{\theta}_j\| > 60°$):

$$CP_{\texttt{attr}}(\hat{\mathbf{X}}_i, \hat{\mathbf{X}}_j) = \exp\left(-\frac{(\hat{d}_{i,j} - ID)^2}{\sigma_{f_2}^2}\right). \tag{18}$$

A static pedestrian can move apart, letting others to pass. This behavior is modeled with a repulsion effect:

$$CP_{\texttt{rep}}(\mathbf{X}_i, \mathbf{X}_j) = 1 - \exp\left(-\frac{d_{i,j}^2}{\sigma_{f_1}^2}\right), \tag{19}$$

with $\sigma_{f_2} = 1m$. Note that a particle can be in both situations at the same time. Only one social force is applied at a time. The SF for this motion model is:

$$SF_4(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,\varphi_i}) = \prod_{j\in\varphi_i} CP(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}), \tag{20}$$

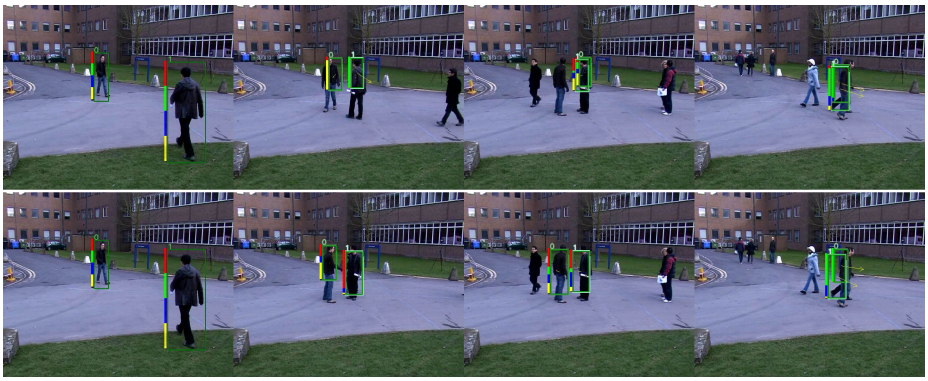

Fig. 3 Example of tracking (central couple only). The top and bottom rows depict the results of our proposal without and with social forces, respectively. We use the view 5 of PETS09 S2-L1 scenario. The rectangles at the left of each bounding box represent the contribution weight of each model. Red for **Stand still**, green for **Going straight**, blue for **Finding one's way** and yellow for **Walking around**.

$$CP(\hat{\mathbf{X}}_i, \hat{\mathbf{X}}_j) = \begin{cases} \texttt{CP}_{attr}(\hat{X}_i, \hat{X}_j) & \text{if } \hat{d}_{i,j} < QD \text{ and} \\ & \hat{\theta}_{i,j} < 60^\circ, \\ \texttt{CP}_{rep}(\hat{X}_i, \hat{X}_j) & \text{if } \hat{d}_{i,j} < QD, \\ 1 & \text{otherwise.} \end{cases}$$

5 Experimental setup, results and evaluation

We have tested our proposal on 6 realistic video sequences to evaluate our results both qualitatively and quantitatively. We have compared our algorithm performance against other proposals from the current state of the art and we show how the social forces model can boost the tracking results.

5.1 Experimental setup

We have used several videos, from three datasets: PETS09 [3], PETS06 [2] and CAVIAR [1]. All datasets give challenging benchmarks to test and evaluate the performance of pedestrian tracking algorithms. The PETS09 dataset consists of a set of 8 camera video sequences of an outdoor scene. We apply our proposal in the sparse crowd scenario S2-L1 (795 frames). The PETS06 dataset is a set of video sequences of an indoor scene from 4 distinct cameras. We use the S6 scenario (2800 frames). Those scenes present challenging situations of pedestrian tracking. Finally, we have also used three sequences from the CAVIAR dataset: EnterExitCrossingPaths1cor (EECP1cor), TwoEnterShop1cor (TES1cor) and TwoLeaveShop2cor (TLS2cor). Those sequences are complementary and cover the situations that can be encountered in this application (occlusion, crowds, interaction, erratic motion, etc.)

We have manually generated a Ground-Truth (GT) dataset, for each pedestrian in the scene over all frames of the views 1 and 2 of the PETS09 S2-L1 scenario and view 4 of the PETS06 S6 scenario. The CAVIAR project provides the GT data. We measure the performance of our algorithm with five standard tracking evaluation metrics [9]: (1) Sequence Frame Detection Accuracy (SFDA) penalizes missed detections and false positive; (2) Average Tracking Accuracy (ATA) penalizes shorter or longer trajectories, missed trajectories and false positive; (3) Multiple Object Tracking Precision (MOTP) and (4) Multiple Object Detection Precision (MODP) measure the tracks spatio-temporal precision and spatial precision, respectively; (5) Multiple Object Detection Accuracy measures the detection accuracy, missed detections and false positives. All those metrics set scores between 0 (worst) and 1 (perfect).

The creation and destruction of the trackers is automatic: From a binary image, coming from a foreground detector algorithm, we initialize new trackers from the detected foreground blobs (regions with motion, see Fig. 2-b), whenever they have the expected dimensions of an adult (with the help of the camera projection matrix, see Fig. 2-c). The tracker is suppressed when its linearized likelihood stays under a threshold for a given time, i.e., 10 frames. The number of particles is fixed initially to 100 for each of the 4 models, so that $N = 400$. This is a compromise between precision (more particles for more precision) and efficiency (more particles mean more computational times). The orientation cue presented in section 3.2 is implemented as in [8], using the same annotated training dataset with 16 image for each one of the 8 discretized directions.

We implemented our algorithm in C++ and we tested it in a PC with an Intel Core i7 processor. Our algorithm allows to process around 5-10 frames per second without special parallelization. This time depends on the number of trackers and on how many of them get into interaction (see Fig. 6), the worst case scenario being when all trackers interact with each other. In this worst case, the SFs have to be computed for all the T trackers with N particles which complexity is $N \cdot T^2$. In our implementation, the orientation estimation is the most time-consuming part since it involves a recurrent computation of HOG feature vectors.

Fig. 4 Example of tracking. Each row depicts the results with the IMM-PF and IMM-PF-SF proposals respectively, using the view 3 of the PETS06 S6 scenario. In the IMM-PF implementation, the tracker 3 switches from one target to another meanwhile in IMM-PF-SF, the identity is preserved. The bounding boxes are the output of our framework where the left rectangles depict the contribution weight of each model. Red for **Stand still**, green for **Going straight**, blue for **Finding one's way** and yellow for **Walking around**.

Sequence	Method	SFDA	ATA	N-MODP	MOTP	MODA
PETS09 View 1	CV	0.67	0.36	0.75	0.73	0.80
	IMM-PF	0.63	0.50	0.77	0.63	0.60
	IMM-PF SF	0.70	0.60	0.82	0.70	0.74
PETS09 View 2	CV	0.51	0.40	0.57	0.56	0.60
	IMM-PF	0.62	0.51	0.85	0.67	0.54
	IMM-PF SF	0.74	0.63	0.91	0.68	0.65
PETS06 View 4	CV	0.33	0.48	0.58	0.50	0.33
	IMM-PF	0.33	0.53	0.66	0.54	0.29
	IMM-PF SF	0.37	0.57	0.73	0.65	0.31
CAVIAR EECP1cor	CV	0.66	0.35	0.88	0.64	0.54
	IMM-PF	0.74	0.63	0.88	0.78	0.68
	IMM-PF SF	0.75	0.67	0.89	0.81	0.68
CAVIAR TES1cor	CV	0.54	0.45	0.77	0.70	0.47
	IMM-PF	0.51	0.57	0.78	0.68	0.30
	IMM-PF SF	0.55	0.59	0.79	0.72	0.40
CAVIAR TLS2cor	CV	0.41	0.29	0.40	0.94	0.34
	IMM-PF	0.54	0.49	0.52	0.82	0.42
	IMM-PF SF	0.53	0.54	0.51	0.87	0.45

Table 1 Results for the six sequences (PETS'09, view 1 and 2, PETS06 and CAVIAR sequences) using: A constant velocity model (CV), our proposal with (IMM-PF SF) and without (IMM-PF) social forces. The median over 30 experiments is shown, with variance inferior to 0.001 in all cases. This proves the approach repeatability, despite the stochastic nature of the particle filter. The best results are in red.

5.2 Results and comparison with other methods

The Figs. 3 and 4 show some qualitative results. The bounding boxes depict the filter output. The rectangles at the left of each bounding box represent the contribution weight of each model, i.e., the dominant color indicates the model that fits best to the dynamic of the target. In these two images, we observe the switch of the motion model. When the target remains in the same position, the dominant color in the left rectangle is red which means that the **Stand still** model is the one who contributes most to the state estimation. When the target moves, the dominant color changes to the associated model whose motion fits best to target speeds.

In Fig. 3, we track only the couple at the center of the image. The top and bottom rows show the tracking results

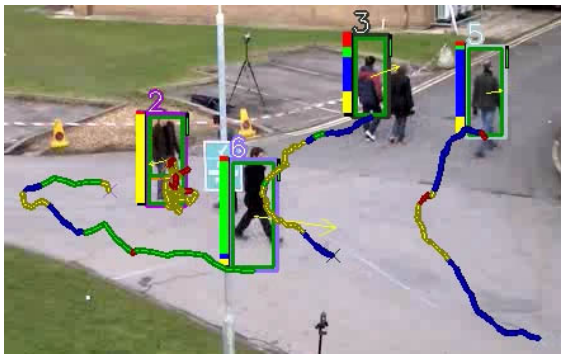

Fig. 5 Tracker trajectories. The lines represent the tracker trajectory for the last 50 frames. The color indicates the model that contributes most to the state estimation. Red is for **Stand still**, green for **Going straight**, blue for **Finding one's way** and yellow for **Walking around**.

with our IMM-PF proposal without and with social forces, respectively. Both targets have similar appearance, hence the trackers on the top (without SF) end following the same target, meanwhile in the bottom row the trackers keep their respective targets. This is due to the repulsion/attraction effect of the **Stand Still** social force model which gives the mayor contribution (i.e., left bar is mostly red in central images). This SF model prevents other tracker particles to follow the same target (repulsion) but also try to keep them at a given distance with opposite orientation (attraction). In this sequence, multiple pedestrians cross in front of the tracked couple. However, our proposed motion model including SFs is robust enough to overcome short partial or total occlusions. The same situation is observed in Fig. 4: the talking couple is correctly tracked meanwhile a tracked pedestrian passes in front and occludes them. The target appearance is kind of similar, especially between tracker 1 and 3, and the pedestrians are moving slowly. In the top row, all trackers end in the same position (one pedestrian is partially occluded by the other) due to the lack of information (appearance/motion). In the bottom row, the couple trackers keep apart by the same phenomenon as in Fig. 3, i.e., the repulsion effect of all SF models aids to preserve the identity of tracker 3. The Fig. 5 depicts the trajectories of the tracker at foot level of the last 50 frames where the color represents the model that contributes more at each frame. One can note that the model switches when there is a change in the trajectory.

In Fig. 6, we depict a representation of the social forces existing between four trackers. The left image is the output of the IMM-PF SF proposal and the right image is the projection of the tracker position in the world plane. In this image, the edges are estimated by the normalized sum of the social forces $SF(\cdot)$ presented in section 4.3. The line thickness is adjusted according to this normalized sum. Thus, the edges only connect those trackers that interact and a thicker line indicates a major influence from that tracker. In this ex-

Fig. 6 Social forces representation. In the left image, we depict the output of our framework with IMM-PF SF. The four trackers are projected to the world plane through camera calibration (right image). The (directed) edges connect the trackers which interact with each other. Edges of same color correspond to the same tracker.

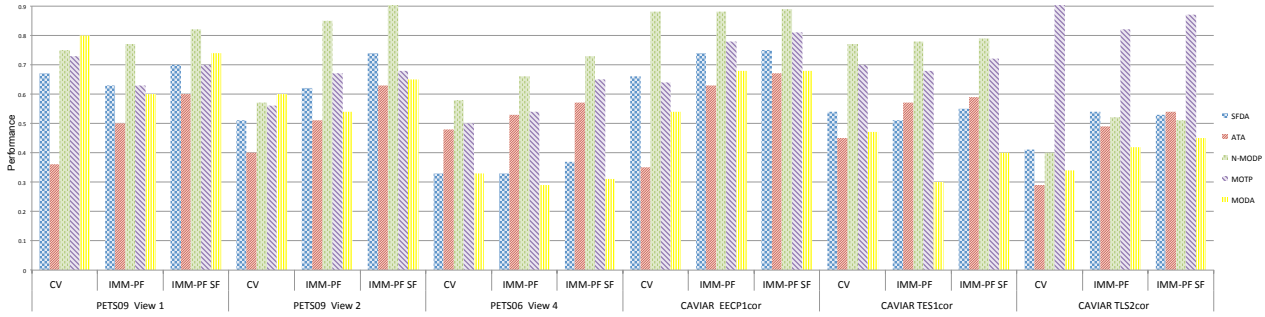

Fig. 7 Results for all sequences (PETS'09, PETS'06 and CAVIAR) using as a motion model: a classic constant velocity model (CV), our proposal with and without including social forces, both with parameter estimations. Median over 30 experiments, with variance inferior to 0.001 in all cases.

ample, tracker 3 is influenced by trackers 1, 2 and 4, while the tracker 0 is far enough not to affect tracker 3.

The table 1 presents quantitative results over the sequence S2-L1 view 1 and 2 of PETS09, view 4 of PETS06 S6 scenario, and the sequences from the CAVIAR dataset. The Fig. 7 depicts a graphical representation of this table. Those are low-density videos with multiple pedestrian interactions (talking people, couple walking). We tested 3 models: a classic constant velocity model (CV), our proposal alone (IMM-PF) and our proposal including the social forces (IMM-PF SF). The rest of the implementation (observation model, initialization, termination, etcetera) remains the same. The SFDA, MODP and MOTP metrics measure the detection precision. In this case, the results show no significant changes for sequences PETS09 View 1, PETS06 View 4 and CAVIAR TES1cor, indicating that our tracking system is robust enough to detect the targets most of the time, under different techniques. On the other hand, we can observe an improvement for the PETS09 View 2 sequence, because the video has multiple occlusions between pedestrians. The MODA metric shows that we can handle correctly the initialization and termination of the trackers. The ATA metric measures the tracking performance. We observe that it is significantly improved with our proposal, meaning that our algorithm can follow a target with the same tracker for more time.

The Fig. 8 compares our best performance (last diagram) against other approaches which were extracted from [9,17]. Once again, our proposal ATA stands out. So, our proposal, with the aide of the SF, can track the same target longer than other techniques that fail preserving the identity of targets with similar appearance. The closest ones are the methods labeled as Yang and Horesh, but it is important to notice that these two approaches perform *multi-camera* tracking, while our system is *monocular*. The SFDA measure (blue column) for Horesh and ours are similar, meaning that both are good enough to detect the pedestrian, minimizing the false positives and missed detections. In this case, Horesh relies on a target detector employed in each frame and we, on the other hand, initialize the tracker by a simple blob detector.

5.3 Discussion

The experimental results show that our method performs well both on indoor and outdoor sequences. The 4 motion cases allow to handle most of the pedestrian dynamics for medium and low dense scenarios. However, our proposal, and more generally any form of tracking with Bayes filters, is not adapted to high-density crowd scenarios, since occlusions may be much longer in that case and most tar-

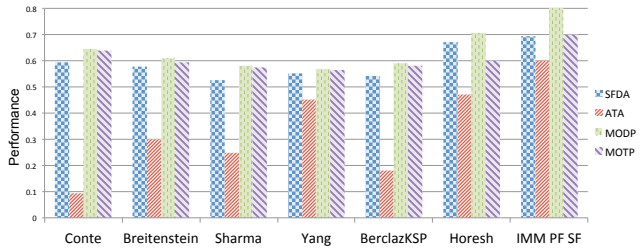

Fig. 8 Evaluation in view 1 of PETS09 S2-L1 sequence. The last diagram shows the performance of our best approach, IMM-PF SF. The others results come from [9,17]. The results labeled Conte, Breitenstein and Shama are monocular tracking system, meanwhile Yang, BerclazKSP and Horesh are multi-view.

gets are barely distinguishable. Also, our proposal may fail more frequently when targets move in completely abnormal ways, i.e., with multiple changes of velocity or direction. Finally, from the PETS results of table 1, we observe that the use of the social forces incorporates the intentionality of the pedestrians in such a way that the trackers interact as people would do, improving the tracking performances. From the CAVIAR results of the same table, we can note that the use of SF does not enhance significantly the score, which is because in these sequences, interactions are rather scarce and short in time.

In fact, ideally, our approach should outperform others in sequences for which the context influences the human trajectories. Given this insight, we have shown results on sequences corresponding to several contexts: outdoor, underground hall, etc.. In environments where the targets have erratic motion or no group interaction but passing by, our approach is less suited. To sum it up, we would expect performances depending on the nature of the sequence and its underlying context.

6 Conclusions and perspectives

We have presented a context-based tracker system with a multiple motion model that includes semantic information of pedestrian behavior for monocular multiple target visual tracking. The IMM-PF allows to handle models with different social content, such as grouping or reactive motion for collision avoidance. The social forces model is a simple and at the same time efficient way to deal with semantic information. The combination of multiple interaction allows our proposal to model high-level behaviors in low-density scenes. The experiments depict how our approach manages efficiently challenging situations that could generate identity switching or target loss.

References

1. CAVIAR dataset. http://homepages.inf.ed.ac.uk/rbf/caviar/
2. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS'2006) www.cvg.rdg.ac.uk/pets2006/
3. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS'2009) www.pets2009.net
4. Andrieu, C., Doucet, A., Holenstein, R.: Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(3), 269–342 (2010)
5. Bazzani, L., Murino, V., Cristani, M.: Decentralized particle filter for joint individual-group tracking. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (2012)
6. Boers, Y., Driessen, J.N.: Interacting multiple model particle filter. In: Proc. of IEEE Conf. on Radar Sonar and Navigation (2003)
7. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera. IEEE Trans. on Pattern Analysis and Machine Intelligence **33**(9), 1820–1833
8. Chen, C., Heili, A., Odobez, J.: Combined estimation of location and body pose in surveillance video. In: Proc. of IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance, pp. 5–10 (2011)
9. Ellis, A., Ferryman, J.: PETS2010 and PETS2009 evaluation of results using individual ground truth single views. In: Proc. of EEE Int. Conf. on Advanced Video and Signal Based Surveillance, pp. 135–142 (2010)
10. Ge, W., Collins, R.T., Ruback, R.B.: Vision-Based Analysis of Small Groups in Pedestrian Crowds. IEEE Trans. on Pattern Analysis and Machine Intelligence **34**(5), 1003–1016 (2012)
11. Hall, E.T.: A system for the notation of proxemic behavior. American anthropologist **65**, 1003–1026 (1963)
12. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. In: Physical review E (1995)
13. Ho, T.J., Chen, B.S.: Novel extended Viterbi-based multiple-model algorithms for state estimation of discrete-time systems with Markov jump parameters. IEEE Trans. on Signal Processing **54**(2), 393–404 (2006)
14. Jiang, Z., Huynh, D.Q., Moran, W., Challa, S.: Tracking pedestrians using smoothed colour histograms in an interacting multiple model framework. In: Proc. of IEEE Int. Conf. on Image Processing (2011)
15. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE Trans. on Pattern Analysis and Machine Intelligence **27**(11), 1805–1819 (2005)
16. Kreucher, C., Hero, A., Keith, K.: Multiple model particle filtering for multitarget tracking. In: Proc. of Workshop on Adaptive Sensor Array Processing (2004)
17. Madrigal, F., Hayet, J.B.: Evaluation of multiple motion models for multiple pedestrian visual tracking. In: Proc. of IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (2013)
18. Mazzon, R., Cavallaro, A.: Multi-camera tracking using a Multi-Goal Social Force Model. Neurocomputing **100**(c), 41–50 (2013)
19. Okamoto, K., Utsumi, A., Ikeda, T., Yamazoe, H., Miyashita, T., Abe, S., Takahashi, K., Hagita, N.: Classification of pedestrian behavior in a shopping mall based on LRF and camera observations. Machine Vision Applications pp. 233–238 (2011)
20. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: IEEE Int. Conf. on Computer Vision, pp. 261–268 (2009)
21. Perez, P., Vermaak, J., Blake, A.: Data fusion for visual tracking with particles. Proc. of the IEEE **92**(3), 495–513 (2004)
22. Shao, J., Jia, Z., Li, Z., Liu, F., Zhao, J., Peng, P.Y.: Spatiotemporal energy modeling for foreground segmentation in multiple object tracking. In: Proc. of the IEEE Int. Conf. on Robotics and Automation (2011)
23. Zhang, S., Das, A., Ding, C., Roy-Chowdhury, A.: Online Social Behavior Modeling for Multi-target Tracking. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshops (2013)