2014 22nd International Conference on Pattern Recognition

Intention-aware multiple pedestrian tracking

Francisco Madrigal*, Jean-Bernard Hayet* and Frédéric Lerasle†
*Centro de Investigación en Matemáticas (CIMAT), Guanajuato, Gto., México
† CNRS, LAAS 7 avenue Colonel Roche, F-31400 Toulouse France
† Univ. de Toulouse, UPS, LAAS, F31400 Toulouse, France
* {pacomd,jbhayet}@cimat.mx †lerasle@laas.fr

***Abstract*—Even though pedestrian motion may look chaotic in most of the cases, recent studies have shown that this motion is mainly ruled by environment and social aspects. In this paper, we propose an interacting multiple model pedestrian tracking framework that incorporates these semantic considerations as a prior knowledge about intentions and interactions between targets. We consider 4 cases of motion for pedestrians: going straight; finding one's way; walking around and standing still. Those models are competing within an Interacting Multiple Model Particle Filter strategy. Targets interactions are handled with social forces, included as potential functions in the weighting process of the Particle Filter (PF). We use different social force models in each motion model to handle high level behaviors (collision avoidance, flocking…). We evaluate our algorithm on challenging datasets and demonstrate that such semantic information improves the tracker performance.**

I. Introduction

Multi-object tracking (MOT) has focused many research efforts in recent years, and is applicable in many areas, like robotics, video surveillance, among others. Among most MOT techniques, stochastic filtering-based approaches infer targets trajectories from two clearly separated elements. The first one is a probabilistic target appearance, and the second one is a probabilistic prior knowledge about the targets motion. Our work focuses on the latter. Pedestrian motion may look chaotic. However, studies [1]–[3] have shown that pedestrian behavior is governed by the pedestrian context, like social forces or environment constraints. For example in Fig. 1, the couple at the center is standing in place, while other people are moving around, in groups or alone, with different velocities. In some cases, as pedestrians in a group, the motion of social targets depends on how they interact with the others and on the intentionality of the interaction. The targets global position and orientation are sufficient to characterize this behavior i.e., to encode information such that "pedestrians in the same group should have similar orientations", or "two nearby people talking to each other should have close to opposite orientations". Such targets interactions are not explicitly considered in most of the approaches that rely in more "naive" dynamic models and independent trackers (i.e., constant velocity [4], [5]). Our claim is that a tracking system modeling these interactions with a semantic dynamic model could dramatically improve the tracking performance. To model these complex dynamics, we simplify our study to four cases of motions (with one model per motion), obtained from a recent analysis of the pedestrians behavior in a mall [2]. Also, we include target interactions (social forces) by using potential functions. The motion models are combined into one single framework with the Interacting Multiple Model (IMM) scheme under a Particle Filter (PF) methodology [6]. In the work we propose here, probabilistic motion models are developed with semantic information from [2], allowing to handle in a more natural way the human walking in sparsely crowded scenes. Our tracking framework dedicates one filter to each target and puts the different motion models in competition. The trackers share semantic information through a prior knowledge of the expected social behavior in each motion model. The modeling considers the body pose of each target (in the vein of [4]) as a feature to control the interaction. We demonstrate that our proposal outperforms existing approaches thanks to large scale comparative evaluations.

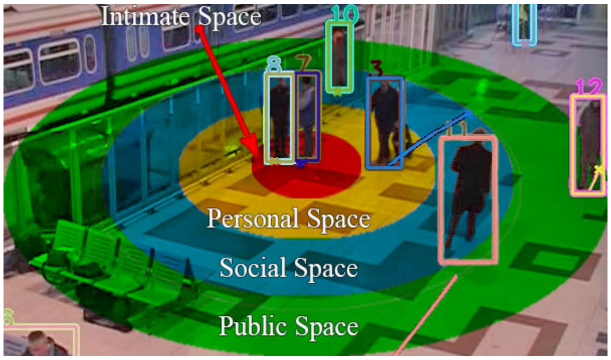

Fig. 1. Pedestrians with multiple motion dynamics. The interaction of the person in the middle of the image with others depends on the region that they occupy. From proxemic theory, these regions can be divided in four: Intimate (Red), Personal (Yellow), Social (Blue) and Public (Green) spaces.

The structure of the paper is as follows: Section II discusses related work. The formulation of IMM-PF is presented in Section III. The Section IV describes our contribution in the modeling of the pedestrian behavior (motion and interaction). Results are presented in Section V. Finally, conclusions are drawn in Section VI.

II. Related work

Naive dynamic models are used in most of the MOT frameworks, i.e., constant velocity model [4], [5], random walks [3], target detector output [5], among others. Unfortunately, those models are only 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. [3] proposes a technique

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.706

to model a simple interaction between trackers. They use a potential function to give more weight to those particles of a PF that are far from other trackers, helping to keep them apart. However, this method can not be extended very well to multiple behaviors since the interaction models can 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, [7] makes use of semantic information to identify groups from independent trackers. [8] solves the tracklet data association problem as a directed graph, by weighting edges according to some social conditions. In [9], the targets interact in such a way that they choose a collision-free 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.

Capturing the complex behavior of targets like pedestrians can be really challenging. Mixing multiple motion models through the IMM methodology is an elegant solution. IMM weights each model according to its importance in the posterior distribution [6], [10]. In [10], target tracking is simulated with a bank of Kalman Filters (KF), where each filter is associated to a distinct linear motion model, within the IMM methodology (IMM-KF). However, the KF cannot use non-linear models and the IMM-KF scheme can not recover when one filter of the bank fails. [6] proposes an IMM implementation with Particle Filter (IMM-PF). They associate a fixed number of particles to each model and weight the models according to their importance in the PF. This proposal suffers from a waste of computational resources when processing many particles with low importance models. In [11], each particle motion model has the possibility of evolving over time, passing from a *moving* to a *stopped* state. Those changes are modeled with a transition matrix (TM) with fixed entries. However, this matrix is a too rough approximation of how the real model changes. On the other hand, target interactions are common in MOT, and the orientation is strongly correlated to the behavior type, i.e., pedestrians from the same group share similar orientations.

Contributions. To overcome the limitations of the common naive dynamic models ([3], [12], [13]), we propose a motion model that introduces semantic information, i.e. some form of intention by the tracked agents. We model this high level pedestrian behavior at two levels: motion and interaction. We emulate the complex pedestrian motion with multiple models, developed from observation analysis [2]. We expand the work of Khan [3] to multiple pedestrian tracking and include more realistic interaction coming from the simulation community, known as social forces. We demonstrate, in several challenging video sequences, that such semantic information improves the tracking performances compared to conventional approaches.

III. Particle Filter-Interacting Multiple Models

The tracking problem is formulated as follows. We infer the state $\mathbf{X}$ in the current time t ($\mathbf{X}_t$) given the set of observations $\mathbf{Z}_{1:t} \overset{\text{def}}{=} \{\mathbf{Z}_1 \dots \mathbf{Z}_t\}$. Under the Markov assumption, the posterior is estimated recursively by Bayesian inference:

$$\left\{ \begin{array}{lcc} p(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) & = & \int p(\mathbf{X}_t|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})d\mathbf{X}_{t-1}, \\ p(\mathbf{X}_t|\mathbf{Z}_{1:t}) & \propto & p(\mathbf{Z}_t|\mathbf{X}_t)p(\mathbf{X}_t|\mathbf{Z}_{1:t-1}). \end{array} \right. \tag{1}$$

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

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

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

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

Since the contribution term does not depend on the previous state $\mathbf{X}_{t-1}$, we move this term out of the integral, leading to

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

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 the models 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 PF approximates the posterior in Eq. 4 by a set of N weighted samples or particles. In addition, we assign one motion model to each particle, as a label $l \in \{1 \dots M\}$. Thereby, a particle n at time t is represented by $(\mathbf{X}_t^{(n)}, \omega_t^{(n)}, l^{(n)})$, where $\omega^{(n)}$ is the particle weight. Each model $m \in \{1 \dots M\}$ has N_m particles associated to it, with a total of $N = \sum_{m=1}^{M} N_m$ particles. The posterior is then approximated by considering both particles weights ($\omega_t^{(n)}$) and models weights (π_t^m):

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

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

A. State definition and proposal distribution

The target state is defined as a Bounding Box (BB) including the target position in the image plane (x, y), its shoulders absolute orientation θ (the angle with respect to the image horizontal line), 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 BB 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. Under the PF scheme, we use an importance proposal distribution $q(\cdot)$, that approximates $p(\mathbf{X}_t|\mathbf{X}_{t-1}, \mathbf{Z}_{1:t})$, 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 the superposition of a deterministic function of the previous state plus a noise sampled from a Gaussian distribution, i.e., $\mathbf{X}_t^{(n)} \sim N(\mathbf{X}_t; tr_{l^{(n)}}(\mathbf{X}_{t-1}^{(n)}), \Sigma_{l^{(n)}})$, where $tr_{l^{(n)}}(\cdot)$ is the deterministic form of the motion model. The index $l^{(n)}$ indicates the model the particle n follows.

B. Observation model and correction step

We implement a probabilistic observation model $p(\mathbf{Z}_t|\mathbf{X}_t)$ based on the proposals presented in [12] and [4]. [12] relies on HSV-space color and motion histograms. We define a reference histogram h_{ref} anytime we create a new tracker. Both likelihoods (based on color $p_c(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$ and motion $p_m(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$, respectively) are evaluated with the Bhattacharya distance between the reference h_{ref} and the current histogram $h^{(n)}$ (corresponding to $\mathbf{X}_t^{(n)}$). We include spatial information with the color observation by using two vertical histograms per target, one for the top part and another for the bottom part.

Following [4], we also include observations related to the target orientation, discretized into eight directions. The body pose angle likelihood is evaluated with a set of multi-level Histogram of Oriented Gradients (HoG) features $f^{(n)}$ extracted from the image, inside each $\mathbf{X}_t^{(n)}$. The idea is to decompose the observed features $f^{(n)}$ as a positive linear combination of a set of training samples for which the orientation has been manually labelled. Then, the orientation likelihood $p_\theta(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$ is calculated as the normalized sum of the weights of the training samples in the determined linear decomposition that share the same (discretized) orientation $\theta_t^{(n)}$ as the particle n. Assuming model independence, the observation model is estimated as $p(\mathbf{Z}_t|\mathbf{X}_t^{(n)}) = p_c(\mathbf{Z}_t|\mathbf{X}_t^{(n)})p_m(\mathbf{Z}_t|\mathbf{X}_t^{(n)})p_\theta(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$, where $p_c(\cdot|\cdot)$ and $p_m(\cdot|\cdot)$ are the color and motion cues proposed in [12] and $p_\theta(\cdot|\cdot)$ is the orientation cue described above [4]. Following the PF scheme, the particles weights are updated by the following expressions:

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

Assuming that the proposal and motion prior distribution are the same, we have:

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

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

Note that Eqs. 6 and 8 ensure that the normalization constraints on Eq. 5 are always satisfied.

C. Resampling

We implement the resampling process as in [14]. It acts in one of two ways:

1.- A sampling done 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 done on a per model basis. Each model has always a minimum of $\gamma = 0.1 * N$ particles to preserve diversity. If the model has less particles than a threshold ($N_m < \gamma$), we draw new samples from a Gaussian distribution: $N(\bar{\mathbf{X}}_{t-1}, \mathbf{S}_{t-1})$, 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 model with more particles to leave the number of particles N unchanged. This resampling manages the model transition implicitly, so no prior transition information is required.

IV. Pedestrian semantic behavior

This section describes our contribution with more details. We propose a motion model for pedestrian tracking that incorporates semantic information about the dynamics of the targets, with a set of expected behavioral rules relying on the concept of interpersonal space between targets (see Fig. 1). In some way, we encode the *motion intentions* in the filter.

A. Priors on pedestrian dynamics

According to [2] there are four pedestrian motion behaviors in a shopping mall:

Going straight. The pedestrians walk 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, 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 transition model:

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

where σ_x, σ_y and σ_θ are constant values. The new position is updated as a constant velocity non-holonomic motion model (known as the unicycle model in the robotics literature). Normally, a pedestrian who walks fast has a rather constant orientation. Following this idea, we calculate the new orientation and angular velocity considering a change in the level of noise, controlled by $\alpha(v) = \exp(\frac{-v^2}{\sigma_\alpha})$. Hence, the higher the linear velocity v_l, the smaller will be the additive 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. In our case, these values are fixed empirically. The **stand still** case is modeled more simply by:

$$tr_4(\mathbf{X}_t) = \begin{bmatrix} I_{3\times3} & 0_{3\times2} \\ 0_{2\times3} & 0_{2\times2} \end{bmatrix} \mathbf{X}_t + \nu_4. \tag{9}$$

where ν_4 is a Gaussian noise. Pedestrians are also influenced by a set of external rules known as social forces (SF) [1]. Those SF depend on the dynamics of the people. The next section describes them in detail.

B. Social behaviors for trackers interaction

The social forces (SF) model makes possible to model the interaction between trackers, and the handling of intention-aware motion models as described before allows us to modulate the interactions based on each target supposed intention. We associate a set of SF to each motion model according to the expected behavior in each case. These behaviors are selected from the proxemics theory [15] and depend on the space occupied by the interacting trackers. In Fig. 1, we depict an example, where the central pedestrian (labeled 2) interacts with the others 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 SF: (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 according to each target motion model, we can cope with many kinds of behaviors. Interactions are modeled with pairwise potential functions [3]. 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 $\varphi_i = \{j \in \{1 \dots N\} : i \neq j\}$. 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 integral with π_t^m. Thus, the posterior of Eq. 4 for a target i is reformulated as:

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

Since the interaction term is out of the predictive part, 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 $\mathbf{x}_{t,i}^{(n)}$ is the particle n state of tracker i and $\hat{\mathbf{x}}_{t,j} = \left[\hat{x}, \hat{y}, \hat{\theta}, \hat{v}_l, \hat{v}_\theta\right]_t^T$ is the weighted mean of all the particles state of tracker j, projected on ground plane (the image-to-scene homography being known), $\hat{r} = [\hat{x}, \hat{y}]^T$ is the position. $SF_{l_i^{(n)}}(\cdot,\cdot)$ is the social force model the particle n is associated to. Our social forces terms are based on the distance between two trackers. We evaluate them through the L2 norm as $\hat{d}_{i,j} = \|\hat{r}_{i,t} - \hat{r}_{j,t}\|$. All the distance considerations in this paper come from the study of nonverbal communication known as proxemics and try to emulate the space depicted in Fig. 1 [15]. We define the social forces for each motion model as follows.

Going straight. The pedestrians who walk straight and fast are aware of the obstacles present in their public space (green circle in Fig. 1) and take decisions with enough anticipation for their direction to ensure a comfortable collision-free path. In that case, we use a repulsion function applying on the tracker from any other tracker under a public distance (green circle in Fig. 1, $PD = 3.5m$): $\hat{d}_{ij} < PD$ (considering a variance of $\sigma_{f_1} = 2m$). The social force for case 1 (sec. IV-A) is:

$$SF_1(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,\varphi_i}) = \prod_{j \in \varphi_i} GS(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}) \tag{10}$$

$$GS(X_i, X_j) = \begin{cases} 1 - \exp\left(-\frac{d_{i,j}^2}{\sigma_{f_1}^2}\right) & \texttt{if } \hat{d}_{i,j} < 3.5m \\ 1 & \texttt{otherwise}. \end{cases}$$

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 $SD = 2.5m$ (blue circle in Fig. 1). 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. We model this as:

$$FW_{\texttt{attr}}(X_i, X_j) = \exp\left(-\frac{(\hat{d}_{i,j} - SD)^2}{\sigma_{f_2}^2}\right). \tag{11}$$

where $\sigma_{f_2}^2 = 20cm$ is the variance over the distance. Otherwise, the target i will tend to evade targets j:

$$FW_{\texttt{rep}}(X_i, X_j) = 1 - \exp\left(-\frac{d_{i,j}^2}{\sigma_{f_3}^2}\right), \tag{12}$$

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

$$SF_2(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,\varphi_i}) = \prod_{j \in \varphi_i} FW(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}) \tag{13}$$

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

Walking around. Pedestrians tend to walk at comfortable speed, in groups. Targets belong to the same group if $\hat{d}_{i,j} < SD$ (the yellow region in Fig. 1), keeping a personal distance of $QD = 1.5m$, 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:

$$WA_{\texttt{attr}}(X_i, X_j) = \exp\left(-\frac{(\hat{d}_{i,j} - QD)^2}{\sigma_{f_2}^2}\right). \tag{14}$$

Otherwise it avoids the obstacles:

$$WA_{\texttt{rep}}(X_i, X_j) = 1 - \exp\left(-\frac{d_{i,j}^2}{\sigma_{f_4}^2}\right), \tag{15}$$

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

$$SF_3(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,\varphi_i}) = \prod_{j \in \varphi_i} WA(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}) \tag{16}$$

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

Stand still. The persons stand still, maybe interacting with other people, i.e., talking, with an interpersonal distance of $ID = 1m$, this is the case in the Fig. 1 where the target 7 speaks with target 8. We model this behavior with an attraction function between two close trackers ($\hat{d}_{i,j} < QD$) with opposite orientations ($\hat{\theta}_{i,j} = \|\hat{\theta}_i - \hat{\theta}_j\| < 60°$):

$$CP_{\texttt{attr}}(\hat{\mathbf{X}}_i, \hat{\mathbf{X}}_j) = \exp\left(-\frac{(\hat{d}_{i,j} - ID)^2}{\sigma_{f_2}^2}\right). \tag{17}$$

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

$$CP_{\texttt{rep}}(X_i, X_j) = 1 - \exp\left(-\frac{d_{i,j}^2}{\sigma_{f_1}^2}\right), \tag{18}$$

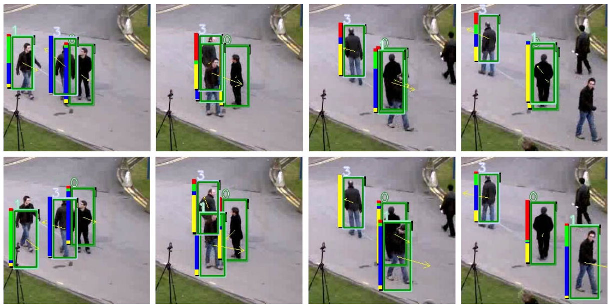

Fig. 2. Example of tracking. The top and bottom rows depict the results of our proposal without and with social forces, respectively. We use the view 1 of PETS09 S2-L1 scenario. The rectangles at the left of each bounding box represent the weight of each model. Red for **Stand still**, green for **Going straight**, blue for **Finding one's way** and yellow for **Walking around**.

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:

$$SF_4(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,\varphi_i}) = \prod_{j \in \varphi_i} CP(\hat{\mathbf{X}}_{t,i}^{(n)}, \hat{\mathbf{X}}_{t,j}) \tag{19}$$

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

V. Experiments

A few more precisions about the tracker implementation need to be detailed: Creating and destroying trackers is done automatically from the binary image resulting from a foreground detector algorithm. New trackers are generated from the detected foreground blobs (regions with motion), only if they have the expected dimensions of an adult (deduced from the projection matrix). The tracker is destroyed when its likelihood stays below a threshold for 10 frames. We evaluated our proposal, both qualitatively and quantitatively, on 3 realistic video sequences, and compare it against other state-of-the-art tracking systems. These videos come from two datasets: PETS09 [16] and PETS06 [17]. Both are challenging benchmark datasets designed explicitly 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. Here, we apply our tracking methodology on two views of the S2-L1 sparse crowd scenario (795 frames each). The PETS06 dataset has a set of 4 camera video sequences of an indoor scene. We use the S6 scenario (2800 frames). Those scenes present challenging situations of pedestrian tracking: occlusions, social interactions, pedestrians with similar appearance…

Sequence	Method	SFDA	ATA	N-MODP	MOTP	MODA
	CV	0.67	0.36	0.75	0.73	0.80
PETS09	IMM-PF	0.63	0.50	0.77	0.63	0.60
View 1	IMM-PF SF	0.69	0.60	0.78	0.68	0.78
	CV	0.51	0.40	0.57	0.56	0.60
PETS09	IMM-PF	0.62	0.51	0.85	0.67	0.54
View 2	IMM-PF SF	0.65	0.59	0.85	0.67	0.61
	CV	0.33	0.48	0.58	0.50	0.33
PETS06	IMM-PF	0.33	0.53	0.66	0.54	0.29
View 4	IMM-PF SF	0.35	0.58	0.68	0.58	0.32

TABLE I. Results for the S2.L1 sequence, view 1. Median over 30 experiments, with variance inferior to 0.001 in all cases. The best approach is in red.

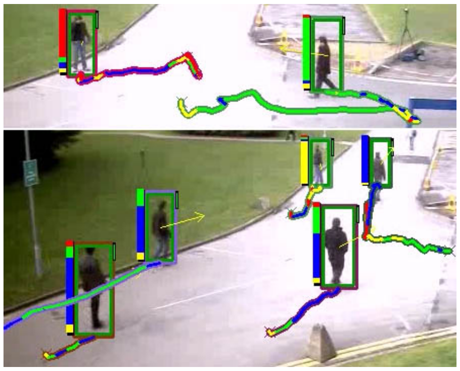

Fig. 3. Tracker trajectories. The lines represent the trajectory of the tracker in the last 70 frames. The color indicates the model that contributes most to the state estimation. Red for **Stand still**, green for **Going straight**, blue for **Finding one's way** and yellow for **Walking around**. Both figures depict the resulting trajectories in two different instants of view 2 of PETS09 S2-L1 scenario.

We generated a ground-truth dataset for these three videos and labeled each pedestrian in the scene over all frames of views 1 and 2 of the PETS09 S2-L1 scenario and view 4 of PETS06 S6 scenario. We evaluate the tracking performance with five classical evaluation metrics [18]: (1) Sequence Frame Detection Accuracy (SFDA), sensitive to missed detections and false positive; (2) Average Tracking Accuracy (ATA), that favors longer trajectories; (3) Multiple Object Tracking Precision (MOTP) and (4) Multiple Object Detection Precision (MODP), measuring the tracks spatio-temporal and spatial precisions, respectively; (5) Multiple Object Detection Accuracy, sensitive to the accuracy, missed detections and false positives. Their scores vary between 0 (worst) and 1 (perfect). First, the Figs. 2 and 3 illustrate qualitative results. The bounding boxes depict the filter output. In Fig. 2, the top and bottom rows show the tracking results with our IMM-PF proposal without and with social forces, respectively. Observe that the three targets have similar appearance, hence the trackers on the top collapse on the same target, meanwhile on the bottom, the trackers keep their respective targets. The Fig. 3 depicts the trajectories of the tracker at foot level of the last 70 frames. The color represents the model that contributes more to the posterior at each frame. One can note that the winning model switches when there is a change in the trajectory. The Table I presents quantitative results over the PETS2009 S2-L1 sequence (views 1 and 2) and PETS06 S6 sequence (view 4). We evaluated 3 models: A classic constant velocity model (CV), our proposal alone (IMM-PF) and our proposal including the social forces (IMM-PF SF). Excepting the motion model, the rest of the implementation is exactly the same. The SFDA, MODP and MOTP metrics, that evaluate the detection precision, are not significantly different for sequences PETS09 View 1 and PETS06 View 4, indicating that our tracking

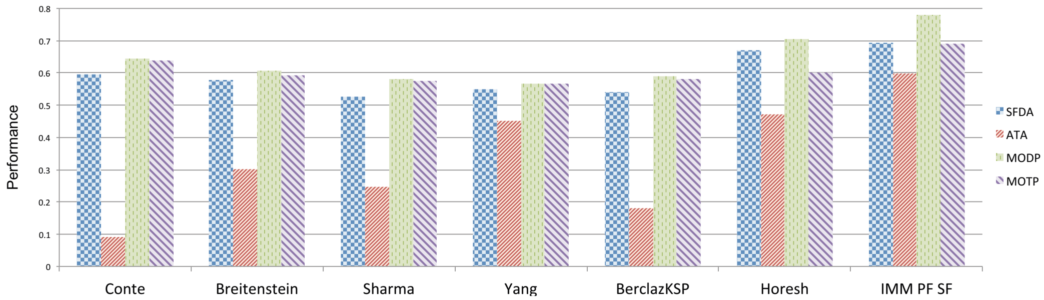

Fig. 4. Evaluation in view 1 of PETS09 S2-L1 sequence. The last diagram shows the performance of our best approach, IMM PF SF. The others results come from [14], [18]. The results labeled Conte, Breitenstein and Shama are monocular tracking system, meanwhile Yang, BerclazKSP and Horesh are multi-view.

system is robust enough to detect the targets most of the time, under different techniques. On the other hand, we observe an improvement on the PETS09 View 2 sequence, because of the presence of multiple occlusions between pedestrians. The MODA fairness shows that we can handle correctly the initialization and termination of the trackers. The ATA metric, measuring the tracking performance, is significantly improved with our proposal, which means that our algorithm can track a given target during longer sequences. Finally, in Fig. 4, we compare our IMM-PF SF proposal (last diagram) against other approaches from the literature, and whose quantitative results on these sequences are available [14], [18]. Once again, the ATA of our proposal stands out. As a consequence, our Social Forces-based proposal can track the same target longer than other techniques, that may fail in preserving the identity of targets with similar appearance. The two methods closest to ours are the ones of Yang and Horesh, but please notice that these 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, and reduce the false positives and missed detections rates. Horesh relies on a target detector employed each frame and we, on the other hand, initialize the tracker by a simple blob detector.

VI. Conclusions

We have presented 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 are a simple and at the same time efficient way to include target interaction. 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] D. Helbing and P. Molnar, “Social force model for pedestrian dynamics,” in *Physical review E*, 1995.

[2] K. Okamoto, A. Utsumi, T. Ikeda, H. Yamazoe, T. Miyashita, S. Abe, K. Takahashi, and N. Hagita, “Classification of pedestrian behavior in a shopping mall based on LRF and camera observations,” *Machine Vision Applications*, pp. 233–238, 2011.

[3] Z. Khan, T. Balch, and F. Dellaert, “Mcmc-based particle filtering for tracking a variable number of interacting targets,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1805–1819, 2005.

[4] C. Chen, A. Heili, and J. Odobez, “Combined estimation of location and body pose in surveillance video,” in *Proc. of IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, 2011, pp. 5–10.

[5] L. Bazzani, V. Murino, and M. Cristani, “Decentralized particle filter for joint individual-group tracking,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[6] Y. Boers and J. N. Driessen, “Interacting multiple model particle filter,” in *Proc. of IEEE Conf. on Radar Sonar and Navigation*, 2003.

[7] W. Ge, R. T. Collins, and R. B. Ruback, “Vision-Based Analysis of Small Groups in Pedestrian Crowds,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1003–1016, May 2012.

[8] S. Zhang, A. Das, C. Ding, and A. Roy-Chowdhury, “Online Social Behavior Modeling for Multi-target Tracking,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2013.

[9] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, “You’ll never walk alone: Modeling social behavior for multi-target tracking,” in *Proc. of Int. Conf. on Computer Vision (ICCV)*, 2009, pp. 261–268.

[10] T.-J. Ho and B.-S. Chen, “Novel extended Viterbi-based multiple-model algorithms for state estimation of discrete-time systems with Markov jump parameters,” *IEEE Trans. on Signal Processing*, vol. 54, no. 2, pp. 393–404, 2006.

[11] C. Kreucher, A. Hero, and K. Keith, “Multiple model particle filtering for multitarget tracking,” in *Proc. of Workshop on Adaptive Sensor Array Processing*, 2004.

[12] P. Perez, J. Vermaak, and A. Blake, “Data fusion for visual tracking with particles,” *Proc. of the IEEE*, vol. 92, no. 3, pp. 495–513, 2004.

[13] Z. Jiang, D. Q. Huynh, W. Moran, and S. Challa, “Tracking pedestrians using smoothed colour histograms in an interacting multiple model framework,” in *Proc. of IEEE Int. Conf. on Image Processing*, 2011.

[14] F. Madrigal and J.-B. Hayet, “Evaluation of multiple motion models for multiple pedestrian visual tracking,” in *Proc. of IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, 2013.

[15] E. T. Hall, “A system for the notation of proxemic behavior,” *American anthropologist*, vol. 65, pp. 1003–1026, 1963.

[16] IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS’2009) www.pets2009.net.

[17] IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS’2006) www.cvg.rdg.ac.uk/pets2006/.

[18] A. Ellis and J. Ferryman, “PETS2010 and PETS2009 evaluation of results using individual ground truth single views,” in *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, 2010, pp. 135–142.