Improvement of multiple pedestrians tracking thanks to semantic information

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

Improvement of multiple pedestrians tracking thanks to semantic information

Francisco Madrigal
Centro de Investigación en Matemáticas (CIMAT)
Guanajuato, Gto., México
Email: pacomd@cimat.mx

Jean-Bernard Hayet
Centro de Investigación en Matemáticas (CIMAT)
Guanajuato, Gto., México
Email: jbhayet@cimat.mx

Frédéric Lerasle
Laboratoire d'Analyse et d'Architecture du Systèmes (LAAS-CNRS)
Toulouse, France
Email: lerasle@laas.fr

***Abstract*—This work presents an interacting multiple pedestrian tracking method for monocular systems that incorporates a prior knowledge about the movement and interactions of the targets. We consider 4 cases of pedestrian behaviors: going straight; finding the way; walking around and stand still. Those 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 (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 improve the tracker performance.**

I. Introduction

Multi-object tracking (MOT) has been the attention of many research efforts in recent years, and is applicable in many areas, like robotics, video surveillance, among others. Among all the MOT techniques, many infer targets trajectories from two clearly separated elements. The first one is the target appearance, and the second one is a prior knowledge about the targets motion. Our work focuses on the latter.

Pedestrian motion may look chaotic. However, many studies [7], [13], [10] have shown that pedestrian behavior is governed by causes such as 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, with different velocities. This complex joint dynamics is not considered in most of the approaches, who rely on more classic individual linear models (i.e., constant velocity). To model these complex dynamics, we simplify the study to four cases of motions (one model per motion), obtained by the analysis of the pedestrians in a mall [13]. Also, we include target interactions by using potential functions, related to the well known social forces. The interaction depends of the orientation of the target, i.e., pedestrians in the same group should have similar orientations, whereas two people talking to each other should have close to opposite orientations. The motion models are integrated in one single framework with the Interacting Multiple Model (IMM) scheme under a Particle Filter (PF) methodology [4].

In the work presented here, the motion models are developed with semantic information from [13], allowing to handle a more natural human walking in low-crowd scenes. Thus, we propose a tracking framework, with a filter dedicated to each target, that includes a prior knowledge of the expected social behavior in each motion model. The modeling considers the body pose of each target (estimated as in[5]) as a feature to control the interaction.

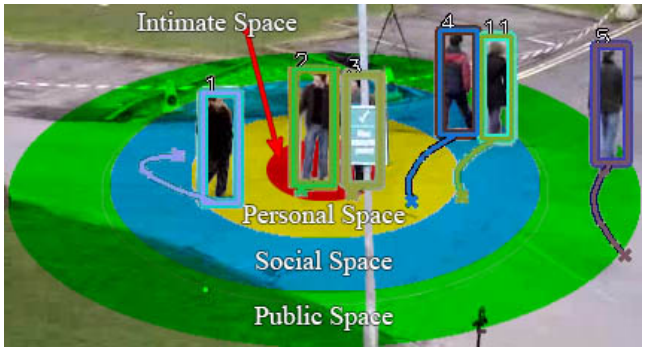

Fig. 1. Pedestrians with multiple motion dynamics. The interaction of the person in the middle of the image with others depends on the region that they occupy. From proxemic theory, these regions can be divided in four: Intimate (Red), Personal (Yellow), Social (Blue) and Public (Green) space.

The 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

Common naive dynamic model are widely used in most of the MOT frameworks, i.e., constant velocity [5], [3], random walk [10], target detector [3], among others. [10] propose a technique to model a simple interaction between trackers to keep them apart. However, this method can not be extended very well to multiple behaviors due to the interaction models can contradict each others. In [3] is presented a framework to track individuals and groups of pedestrians at the same time, but it not consider semantic information over the group formation. Capturing the complex behavior of targets like pedestrians can be really challenging. A mixture of multiple motion models through IMM methodology is an elegant solution. IMM weights each model according to its importance in the posterior distribution [4], [8], fusing models of different types under one single context. In [8], a simulation of target tracking is done with a bank of Kalman Filters (KF), each one

associated with a distinct linear motion model, within the IMM methodology (IMM-KF). This proposal is fast and suitable for a large number of targets. In [15], 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 foreground subtraction. [9] proposes a IMM-KF for pedestrian tracking similar to ours, with two classic motion models: constant position and constant velocity, to track a few targets.

However, the limitations of the KF make impossible to use them in non-linear models and the IMM-KF scheme can not recover when one filter of the bank fails. [4] proposes an IMM implemented with Particle Filter (IMM-PF). They associate to each model a fixed number of particles (i.e., 1000) and weight the models according to their importance in the PF. This proposal suffers from a waste of computational resources at processing a bunch of particles with low importance models. In [11], each particle motion model has the possibility of evolve over time, passing from a *moving* to a *stopped* state. Those changes are modeled with a transition matrix (TM) of fixed values. However, those values can not represent the real model changes. In the other hand, target interaction are common in MOT, and the orientation is strongly correlated to the behavior type, i.e., pedestrians from the same group share similar orientations. [10] presents a strategy to model interactions as potential functions easily included in the Bayes filter formulation. They follow a PF strategy, where the interaction functions act as weight factors in the particle weighting.

Contributions. To overcome the limitation of the common naive dynamic models (widely used in MOT [14], [9], [10]), we propose a motion model incorporating semantic information to improve pedestrian tracking. We model this high level pedestrian behavior in two contexts: motion and interaction. We emulate the complex pedestrian motion with multiple models, developed from observation analysis [13]. We expand the work of Khan [10] 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 the semantic information improves tracking performances.

III. Particle Filter-Interacting Multiple Models

We formulate the tracking problem in a Bayesian framework, where we infer the state $\mathbf{X}$ in the current time t ($\mathbf{X}_t$) given the set of observations $Z_{1:t} = \{Z_1 \dots Z_t\}$. Under the Markov assumption, the posterior is estimated recursively by:

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

The Eq. 1 is know as the Bayes filter which includes two steps: prediction (first row) and correction (second row). Following the IMM strategy [4], we formulate the motion model $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ as a mixture of M distribution 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 term 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 [4]: $\pi_t^m = \pi_{t-1}^m \int p(\mathbf{Z}_t|\mathbf{X}_t) p^m(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) d\mathbf{X}_t$.

A. Tracker implementation

The target state is defined through a bounding box (BB) including its position in the image plane (x, y), its global orientation θ (linked to the shoulders) and its linear and angular velocities (v_l, v_θ). Hence, we define the state $\mathbf{X}$ as $(x, y, \theta, v_l, v_\theta)^T$. The real 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. The PF 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, 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):

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

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

B. Sampling and dynamic model

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 of each particle from the motion model corresponding to its label $l^{(n)}$. This model is a Gaussian distribution $N(\mathbf{X}_t; tr_{l^{(n)}}(\mathbf{X}_{t-1}^{(n)}), \Sigma_{l^{(n)}})$, where $tr_{l^{(n)}}(\cdot)$ is the deterministic form of the motion model. The index $l^{(n)}$ indicates the model the particle n follows.

C. Observation model and correction step

We implemented a probabilistic observation model $p(\mathbf{Z}_t|\mathbf{X}_t)$ based on the proposals presented in [14] and [5].

[14] 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 two vertical histograms per target, one for the top part of the person and another for the bottom part.

Following [5], we include observations related to the target orientation, discretized into eight directions. The body pose angle is evaluated with a set of multi-level Histogram of Oriented Gradients (HoG) features $f^{(n)}$extracted from the image inside each $\mathbf{X}_t^{(n)}$, and decomposed into a linear combination of O training samples $\mathbf{F} = \{f_1, \ldots, f_O\}$:

$$f^{(n)} \approx a_1 f_1 + \cdots + a_O f_O = \mathbf{Fa},$$

where $\mathbf{a} = \{a_1, \ldots, a_O\}$ is the weighs vector subject to non-negative constraints ($a_o \geq 0$). The goal is to find the optimal weights ($\mathbf{a}^*$) such as:

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

where λ controls the regularizor importance. The likelihood is calculated as the normalized sum of the weights of $\mathbf{a}^*$ with the same (discretized) orientation $\theta_t^{(n)}$ of the particle n.

Then, particles weights are updated by:

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

where $p(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$ is the likelihood of the observation $\mathbf{Z}_t$ evaluated at the state of particle n. Assuming that the proposal and prior distribution are the same, we have:

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

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

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

D. Resampling

We implement the resampling process proposed in [12] which performs the sampling in one of two ways:

1.- The sampling is 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 to the target motion.

2.- The resampling is 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

We propose a motion model for pedestrian tracking that incorporate semantic information of the dynamic of the targets with a set of expected behavioral rules in each case.

A. Pedestrian dynamics

According to [13] 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 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 could be interacting with other persons.

We build 4 motion models to emulate those behaviors. The first three cases ($k = 1, 2, 3$) are modeled by:

$$tr_k(\mathbf{X}) = \begin{bmatrix} x + v_l * \cos(v_\theta) \\ y + v_l * \sin(v_\theta) \\ \alpha(v_l) * \theta + (1 - \alpha(v_l)) * v_\theta \\ \mu_k \\ v_\theta \end{bmatrix} + \begin{bmatrix} N(0, \sigma_x) \\ N(0, \sigma_y) \\ N(0, \sigma_\theta) \\ N(0, \sigma_{v_{l,k}}) \\ N(0, \sigma_{v_{\theta,k}}) \end{bmatrix},$$

where σ_x, σ_y and σ_θ are constant values. The new position is updated as a constant velocity model. Normally, a pedestrian who walks fast has a rather constant orientation and small angular velocity. Following this idea, we calculate the new orientation as a combination of the previous angular velocity and orientation, controlled by $\alpha(v) = \exp(\frac{-v^2}{\sigma_\alpha})$. Hence, the higher v_l, the more similar the orientation and angular velocities. The μ_k and σ_k values depend on the model to be used, allowing to control the behavior of the aforementioned categories 1, 2 and 3:

$$\begin{array}{lll} \mu_1 = 2.5\frac{m}{s}, & \sigma_{v_{l,1}} = 0.22\frac{m^2}{s^2}, & \sigma_{v_{\theta,1}} = 0.01\frac{\text{rad}}{s}, \\ \mu_2 = 1.25\frac{m}{s}, & \sigma_{v_{l,2}} = 0.58\frac{m^2}{s^2}, & \sigma_{v_{\theta,2}} = 0.05\frac{\text{rad}}{s}, \\ \mu_3 = 0.65\frac{m}{s}, & \sigma_{v_{l,3}} = 0.63\frac{m^2}{s^2}, & \sigma_{v_{\theta,3}} = 0.10\frac{\text{rad}}{s}. \end{array}$$

The **stand still** case is modeled 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) [7]. Those SF depend on the dynamic of the people. The next section describes them in detail.

B. Social behaviors for trackers interaction

The social forces (SF) model makes possible the interaction between trackers. We associate a set of SF to each motion model according to the expected behavior in each case. 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 of each other, preventing identity switching and collisions. (2) An attraction force, which keeps the targets close to each other, and modeling social groups.

Interactions are modeled with pairwise potential functions [10]. 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 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 happened in Eq. 3, the interaction term $SF_m(\cdot)$ does not depend on the previous state $\mathbf{X}_{t-1}$, so, this term is move out of the mixture distribution with π_t^m. Thus, the posterior of Eq. 4 for a target i is reformulated as:

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

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 ground plane through the homography, $\hat{r} = [\hat{x}, \hat{y}]^T$ is the position, $\hat{\theta}$ is the orientation and $(\hat{v}_l, \hat{v}_\theta)$ is the linear and angular velocity of the target in world coordinates. The term $SF_{l_i^{(n)}}(\cdot,\cdot)$ is the corresponding 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 (see Fig. 1). We define the social forces for each motion models as:

Going straight. The pedestrians who walk fast are aware of the present obstacles and decide with enough anticipation their direction to ensure a comfortable free-collision path. We use a repulsion function over any tracker under a distance of $\hat{d}_{ij} < 3.5m$ (considering a variance of $\sigma_{f_1} = 2m$). The social force for case 1 (sec. IV-A) is:

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

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

Finding one's way. The pedestrian walks at middle/high speed, moving alone, inside a group or merges/splits from one. At this speed, groups are not too close, preserving a social distance $SD = 2.5m$. We consider that two targets with $\hat{d}_{i,j} < 3.5$, $\|\hat{v}_{l,i} - \hat{v}_{l,j}\| < \epsilon_v$, and orientation $\|\hat{\theta}_i - \hat{\theta}_j\| < \epsilon_\theta$ are walking in a 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 is moving alone, evading obstacles:

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

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

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

$$FW(X_i, X_j) = \begin{cases} FW_{\texttt{attr}}(X_i, X_j) & \texttt{if } \hat{d}_{i,j} < 3.5m \\ & \quad \|\hat{v}_{l,i} - \hat{v}_{l,j}\| < \epsilon_v \\ & \quad \|\hat{\theta}_i - \hat{\theta}_j\| < \epsilon_\theta \\ FW_{\texttt{rep}}(X_i, X_j) & \texttt{if } \hat{d}_{i,j} < 3.5m \\ 1 & \texttt{otherwise} \end{cases}$$

Walking around. Pedestrians walk with a comfortable speed, in groups. Targets belong to the same group if $\hat{d}_{i,j} < 3$, $\|\hat{v}_{l,i} - \hat{v}_{l,j}\| < \epsilon_v$ and $\|\hat{\theta}_i - \hat{\theta}_j\| < \epsilon_\theta$. Usually, they keep a personal distance of $SP = 1.5m$. This flock behavior is modeled as:

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

Otherwise it is walking alone, avoiding 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:

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

$$WA(X_i, X_j) = \begin{cases} WA_{\texttt{attr}}(X_i, X_j) & \texttt{if } \hat{d}_{i,j} < 3m \\ & \quad \|\hat{v}_{l,i} - \hat{v}_{l,j}\| < \epsilon_v \\ & \quad \|\hat{\theta}_i - \hat{\theta}_j\| < \epsilon_\theta \\ WA_{\texttt{rep}}(X_i, X_j) & \texttt{if } \hat{d}_{i,j} < 3m \\ 1 & \texttt{otherwise} \end{cases}$$

Constant position. The persons stand still, maybe interacting with other people, i.e., talking, with an interpersonal distance

Fig. 2. Tracking of the 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.

of $ID = 1m$. We model this behavior with an attraction function between two close trackers ($\hat{d}_{i,j} < 1.5m$) with opposite orientation ($\hat{\theta}_{i,j} = \|\hat{\theta}_i - \hat{\theta}_j\| < 60°$):

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

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

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

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 the a time. The SF for this motion model is:

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

$$CP(\hat{X}_i, \hat{X}_j) = \begin{cases} \texttt{CP}_{attr}(\hat{X}_i, \hat{X}_j) & \texttt{if } \hat{d}_{i,j} < 1.5m \\ & \hat{\theta}_{i,j} < 60° \\ \texttt{CP}_{rep}(\hat{X}_i, \hat{X}_j) & \texttt{if } \hat{d}_{i,j} < 1.5m \\ 1 & \texttt{otherwise} \end{cases}$$

V. Experiments

We test our proposal on 3 realistic video sequences evaluating our results qualitatively and quantitatively. We compare our algorithm performance against other proposals from the current state of the art.

We evaluate our proposal with several videos from two datasets: PETS09 [2] and PETS06 [1]. Both are challenging benchmark data 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 of 795 frames. The PETS06 dataset has a set of 4 camera video sequences of an indoor scene. We use the S6 scenario (of 2800 frames). Those scenes present challenging situation of pedestrian tracking.

We manually generated a ground-truth 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 PETS06 S6 scenario. We measure the performance of our algorithm with five standard metrics of tracking evaluation [6]: (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. Those metrics set scores between 0 (worst) and 1 (perfect).

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	0.69	0.60	0.78	0.68	0.78
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	0.65	0.59	0.85	0.67	0.61
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	0.35	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 creation and destruction of the trackers is automatic [12]. From a binary image, coming from a foreground detector algorithm, we initialize new trackers from those blobs (region with motion) that have the expected dimensions of a adult human (with the help of the camera projection matrix). The tracker is destroyed when its linearized likelihood is under a threshold for a given time, i.e., 10 frames.

The Figs. 2 and 3 show some qualitative results. The bounding boxes (BB) depict the filter output. In Fig. 2, we track only the couple at the center of the image. The top and bottom rows show the tracking results with our IMM-PF proposal without and with social forces, respectively. Both targets have similar appearance, hence the trackers on the top end following the same target, meanwhile in the bottom row the trackers keep their respective targets. The same situation is observed in Fig. 3: the talking couple is correctly tracked. The last column depicts a pedestrian that passed in front of them. In the social force case, their identity is preserved.

The Table I presents quantitative results over the sequence S2-L1 view 1 and 2 of PETS09 and view 4 of PETS06 S6 scenario. We tested 3 models: a classic constant velocity model (CV), our proposal alone (IMM-PF) and 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 and PETS06 View 4, indicating that our tracking system is robust enough to detect the targets most of the time, under different techniques. In 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 tracking performance. We observe that ATA is significantly improved with our proposal,

Fig. 3. Example of tracking. The upper row implements IMM-PF with the four proposed motion models. The bottom row depicts the result of our proposal with the four models with social forces. We use the view 3 of PETS06 S6 scenario.

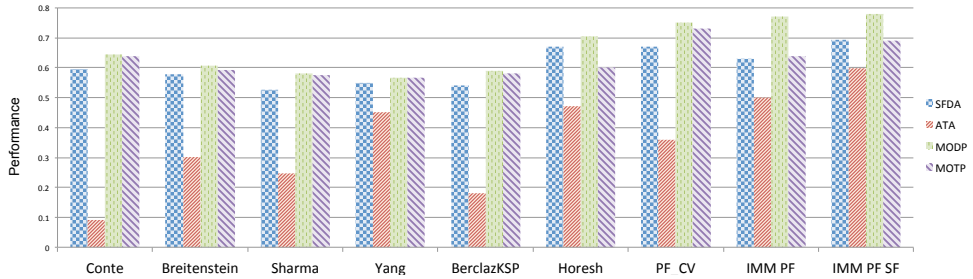

Fig. 4. Evaluation in view 1 of PETS09 S2-L1 sequence. PF_CV follows our implementation with a constant velocity model. The last two results show the performance of IMM PF (no social forces) and IMM PF SF (social forces). The others results come from [6], [12].

meaning that our algorithm can follow a target with the same tracker for more time.

The Fig. 4 compares our performance (last two diagrams) against other approaches. Once again, our proposal ATA stands out. So, our proposal can track the same target longer than other techniques. The two closest ones are the methods labelled as Yang and Horseh, but it is important to notice that these two proposals perform *multi-camera* tracking, while our system is *monocular*.

VI. Conclusions and Future Works

We have presented a multiple motion model that include 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] IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance 2006 www.cvg.rdg.ac.uk/pets2006/.

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

[3] L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[4] Y. Boers and J. N. Driessen. Interacting multiple model particle filter. In *Proc. of the IEEE Conf. on Radar Sonar and Navigation*, 2003.

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

[6] A. Ellis and J. Ferryman. Pets2010 and pets2009 evaluation of results using individual ground truth single views. In *Proc. of the IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, pages 135–142, 2010.

[7] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. In *Physical review E*, 1995.

[8] 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,*, 54(2):393–404, 2006.

[9] Z. Jiang, D. Q. Huynh, W. L. J. Moran, and S. Challa. Tracking pedestrians using smoothed colour histograms in an interacting multiple model framework. In *Proc. of the IEEE Int. Conf. on Image Processing (ICIP)*, 2011.

[10] 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*, 27(11):1805–1819, 2005.

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

[13] 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*, pages 233–238, 2011.

[14] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. of the IEEE*, 92(3):495–513, 2004.

[15] J. Shao, Z. Jia, Z.-p. Li, F.-Q. Liu, and J. Zhao. Spatiotemporal energy modeling for foreground segmentation in multiple object tracking. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2011.