Who are you with and Where are you going?

Kota Yamaguchi Alexander C. Berg Luis E. Ortiz Tamara L. Berg
Stony Brook University
Stony Brook University, NY 11794, USA
{kyamagu, aberg, leortiz, tlberg}@cs.stonybrook.edu

Abstract

We propose an agent-based behavioral model of pedestrians to improve tracking performance in realistic scenarios. In this model, we view pedestrians as decision-making agents who consider a plethora of personal, social, and environmental factors to decide where to go next. We formulate prediction of pedestrian behavior as an energy minimization on this model. Two of our main contributions are simple, yet effective estimates of pedestrian destination and social relationships (groups). Our final contribution is to incorporate these hidden properties into an energy formulation that results in accurate behavioral prediction. We evaluate both our estimates of destination and grouping, as well as our accuracy at prediction and tracking against state of the art behavioral model and show improvements, especially in the challenging observational situation of infrequent appearance observations – something that might occur in thousands of webcams available on the Internet.

1. Introduction

Despite many recent advances in tracking algorithms, effective tracking in realistic scenarios is still quite challenging. One common, yet less well studied scenario, is surveillance of scenes with infrequent appearance observations – such as the sporadic frames one would get from the thousands of webcams streaming pictures from around the globe. In this case, the video stream consists of images that are low resolution, low frame rate (sometimes every few seconds), and display uncontrolled lighting conditions. Additional confusion can result from occlusion between multiple targets due to crowding. Having a strong prior on what we observe will be essential for successful tracking in these challenging situations. In this paper, we look at low frame rate and crowded tracking scenarios with a focus on the behavioral model of pedestrians. This focus helps us both predict where people will go, and who they are with, and leads to improved tracking results.

Pedestrians exhibit complex behavior from various social and environmental factors. For instance, a pedestrian has his or her own destination in mind a comfortable walking speed, and plans a motion path that avoids other pedestrians and scene obstacles. Our goal in this paper is to build a behavioral model that takes into account these higher level decisions and which can easily be "plugged" into existing appearance-based algorithms. With this in mind, we model individual pedestrians as agents who make decisions about velocity in the next time step, given factors from the scene (e.g. other pedestrians to avoid or walk with, or obstacles). We frame this decision process as minimization of an energy function that encodes physical condition, personal motivation, and social interaction.

One aspect of our approach is that we explicitly address the problem of estimating hidden factors that might effect a pedestrian's decision making. One factor is the desired grouping behavior – who a pedestrian is trying to walk with. Another is the pedestrian's desired destination in the scene. Neither of these factors is usually known in the surveillance setting. We estimate these hidden *personal properties* by viewing them as a classification problem, predictable from the trajectory of a pedestrian in the scene. In a surveillance scenario, it is reasonable to assume that there is a set of a few destinations in the scene, such as the entrance of a building. This naturally limits the pattern of trajectories in the scene. Also, people undergoing social interactions tend to show a unique behavioral pattern compared with individuals moving alone. We define a feature representation of trajectories on top of our velocity observations, and predict both of these hidden personal variables using efficient classification approaches.

The contributions of this paper are: 1) producing an explicit energy function based behavioral model that encodes personal, social, and environmental decision factors, 2) data-driven estimation of hidden personal properties that affect the behavior of pedestrians, and 3) use of our proposed behavioral model for improved tracking performance in low frame rate scenarios. We emphasize that our energy function considers social interactions (grouping of pedestrians as they walk, talk and interact), a factor which has only recently been explored in [9]. Our approach to social group estimation is simpler, and more computationally efficient, while remaining effective.

This paper is organized as follows. Section 2 describes related work. Section 3 describes our comprehensive behavioral model, followed by parameter learning in Section 4. Section 5 details the estimation method of hidden personal properties using trajectory features. Section 6 describes the quantitative evaluation of our behavioral model and property estimation with application in tracking, and Section 7 concludes this paper.

2. Related work

The pedestrian behavior model has been extensively studied in the fields where simulation plays an important role, such as graphics [5], and civil engineering [10] [11], or where accurate prediction is required, such as robotics [13, 6]. In most crowd simulation contexts, the base model dates back to the classic social force model [4], in which the behavioral factors are assumed to give an equation that drives pedestrians in analogy to physics. In computer vision, the attempt to detect abnormal events with the social force model is reported in [7].

[12], several social factors are known to affect a person's behavior. Antonini's work [1] is one of the earliest in computer vision to take advantage of the rich behavioral information in a tracking application. The discrete choice model in their work assumes that individual pedestrians make a choice from a discrete set of velocity options at each time step based on social and environmental factors in the scene. The assumption of discretized choice allows efficient prediction with analytical solutions despite the large number of factors considered in the model [1, 10]. However, due to the nature of the discretization, the behavioral prediction tends to show artifacts when metric is continuous. In contrast, continuous models have been recently proposed by [8, 11]. An advantage of continuous model is the flexibility of constructing complex models, however, previous work focuses on individual motivation of the behavior [8, 11], and the only social context is collision avoidance.

of social interaction in behavioral model. Social interaction in the pedestrian group began to be studied only recently in computer vision [2, 3, 9]. A tracking application is included in [9]. There the problem is formulated as a simultaneous discrete assignment of hypothetical tracks and estimation of social relationships based on observations over a short time frame using a CRF. The CRF formulation indirectly encodes a behavioral model. Our focus is to build an explicit behavioral model which can exploit the rich behavioral context in social interactions, yet remain straightforward and efficient enough to be plugged into other tracking approaches as a module.

3. Behavioral model

3.1. An energy function for pedestrian navigation

Our behavioral model is based on an energy function for each pedestrian that expresses the desirability of possible directions of motion for the pedestrian. The energy function combines terms for the various factors that influence the pedestrian's choice. These are explained in this section. We optimize the parameters of this model so that choosing the minimum energy direction accurately predicts the behaviors of pedestrians in labeled training data, and then evaluate the performance of the model on previously unseen test data. The fitting procedure is described in Section 3.2.

At each time step t, pedestrian i is represented by a state variable $s_i^{(t)} = (\mathbf{p}_i^{(t)}, \mathbf{v}_i^{(t)}, u_i^{(t)}, \mathbf{z}_i^{(t)}, A_i^{(t)})$, where $\mathbf{p}_i^{(t)}$, $\mathbf{v}_i^{(t)}$, $u_i^{(t)}$ and $\mathbf{z}_i^{(t)}$ are the position, velocity, preferred speed and chosen destination, respectively, of pedestrian i at time t, while A_i is the set of pedestrians in the same social group as pedestrian i, including himself. Note that $u_i^{(t)}$, $\mathbf{z}_i^{(t)}$ and A_i are not observable and usually assumed static, i.e., $u_i^{(t)} = u_i$, $\mathbf{z}_i^{(t)} = \mathbf{z}_i$ and $A_i^{(t)} = A_i$ are time-invariant[1]. As in [8], our model assumes that each pedestrian makes a decision on the velocity $\mathbf{v}_i^{(t+\Delta t)}$ based on various environmental and social factors in the scene, and we model this decision-making process as the minimization of an energy function.

Our energy function E_Θ, where $\Theta = \{\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \sigma_d, \sigma_w, \beta\}$ denotes a set of parameters, is as follows and consists of a linear combination[2] of six components:

$$\begin{aligned} E_\Theta(\mathbf{v}; s_i, \boldsymbol{s}_{-i}) \equiv & \lambda_0 \; E_{\text{damping}}(\mathbf{v}; s_i) + \\ & \lambda_1 \; E_{\text{speed}}(\mathbf{v}; s_i) + \\ & \lambda_2 \; E_{\text{direction}}(\mathbf{v}; s_i) + \\ & \lambda_3 \; E_{\text{attraction}}(\mathbf{v}; s_i, \boldsymbol{s}_{A_i}) + \\ & \lambda_4 \; E_{\text{group}}(\mathbf{v}; s_i, \boldsymbol{s}_{A_i}) + \\ & E_{\text{collision}}(\mathbf{v}; s_i, \boldsymbol{s}_{-i} \mid \sigma_d, \sigma_w, \beta), \end{aligned} \quad (1)$$

where we define $\boldsymbol{s}_{A_i}$ to be a set of state variables of the pedestrians in i's social group A_i, and $\boldsymbol{s}_{-i}$ to be the set of states of other pedestrians except i. From now on, the time step t is dropped from each variable for notational simplicity.

The following paragraphs provide a description of each of the six components of the energy function E_Θ.

Damping. The damping term penalizes sudden changes in the choice of velocity, relative to the current state:

$$E_{\text{damping}}(\mathbf{v}; s_i) \equiv |\mathbf{v} - \mathbf{v}_i|^2 \ . \quad (2)$$

[1] Sec. 4 shows how we automatically estimate these.

[2] The coefficients are relative, so we fix the collision coefficient to 1.

Speed. Pedestrians have their own preferred speed depending on physical state, culture or scene environment. The speed term penalizes choosing a speed that deviates from the (hidden) preferred speed u_i of the pedestrian i:

$$E_{\text{speed}}(\mathbf{v}; s_i) \equiv (u_i - |\mathbf{v}|)^2 \; . \tag{3}$$

Direction. The direction term concerns the choice of the correct direction towards the goal. We model this by using the negative cosine between the velocity choice $\mathbf{v}$ and the direction to the destination $\mathbf{z}_i$ from the current location $\mathbf{p}_i$:

$$E_{\text{direction}}(\mathbf{v}; s_i) \equiv -\frac{\mathbf{z}_i - \mathbf{p}_i}{|\mathbf{z}_i - \mathbf{p}_i|} \cdot \frac{\mathbf{v}}{|\mathbf{v}|} \; . \tag{4}$$

Attraction. People in the same group tend to stay close to each other while moving together. To capture this effect, we define the attraction term as

$$\begin{aligned} &E_{\text{attraction}}(\mathbf{v}; s_i, \mathbf{s}_{A_i}) \equiv \\ &\textstyle\sum_{j \in A_i - \{i\}} \left(\frac{\mathbf{v}_i}{|\mathbf{v}_i|} \cdot \frac{\mathbf{v}_j}{|\mathbf{v}_j|}\right) \left(\frac{\Delta \mathbf{p}_{ij}}{|\Delta \mathbf{p}_{ij}|} \cdot \frac{\mathbf{v}}{|\mathbf{v}|}\right) \end{aligned} \tag{5}$$

where $\Delta \mathbf{p}_{ij} = \mathbf{p}_i - \mathbf{p}_j$. The second factor penalizes choosing a forward direction that is far from another pedestrian $j \in A_i - \{i\}$ in the group A_i of pedestrian i. The first factor is a weight that flips this attraction effect if person j is moving in a direction opposite to i.

Grouping. People in the same group tend to walk at similar speeds and directions. The grouping term penalizes velocity choices that are different from the average velocity of the group:

$$E_{\text{group}}(\mathbf{v}; s_i, \mathbf{s}_{A_i}) \equiv |\mathbf{v} - \bar{\mathbf{v}}_{A_i}|^2 \tag{6}$$

$$\text{where } \bar{\mathbf{v}}_{A_i} \equiv \tfrac{1}{|A_i|} \textstyle\sum_{j \in A_i} \mathbf{v}_j \; . \tag{7}$$

Note that the social group A_i always includes pedestrian i. If A_i is a singleton, the grouping term has the same effect as the damping term.

Collision. Pedestrians try to avoid collisions with obstacles or other pedestrians. We use the model described in [8] to capture this effect:

$$\begin{aligned} &E_{\text{collision}}(\mathbf{v}; s_i, \mathbf{s}_{-i} \mid \sigma_d, \sigma_w, \beta) \equiv \\ &\textstyle\sum_{j \neq i} w(s_i, s_j) \exp\left(-\frac{d^2(\mathbf{v}, s_i, s_j)}{2\sigma_d^2}\right) \; . \end{aligned} \tag{8}$$

Note that this term requires three parameters σ_d, σ_w, and β. The factor $w(s_i, s_j)$ is a weight coefficient, while the function $d(\mathbf{v}, s_i, s_j)$ in the exponent is the expected minimum distance between pedestrian i and j under a constant-velocity assumption [8]:

$$w(s_i, s_j) \equiv \exp\left(-\frac{|\Delta \mathbf{p}_{ij}|^2}{2\sigma_w^2}\right) \cdot \left(\frac{1}{2}\left(1 - \frac{\Delta \mathbf{p}_{ij}}{|\Delta \mathbf{p}_{ij}|} \cdot \frac{\mathbf{v}_i}{|\mathbf{v}_i|}\right)\right)^{\beta} \tag{9}$$

$$d(\mathbf{v}, s_i, s_j) \equiv \left| \Delta \mathbf{p}_{ij} - \frac{\Delta \mathbf{p}_{ij} \cdot (\mathbf{v} - \mathbf{v}_j)}{|\mathbf{v} - \mathbf{v}_j|^2} (\mathbf{v} - \mathbf{v}_j) \right| \tag{10}$$

The first term in (9) assigns less influence to distant pedestrians, while the second term in (9) assigns less weight to pedestrians outside the view of pedestrian i.

3.2. Dynamical model

We now describe how to fit the parameters of the model. Recall our assumption that the personal properties of individual pedestrians are static, i.e., $u_i^{(t)} \approx u_i$, $\mathbf{z}_i^{(t)} \approx \mathbf{z}_i$, and $A_i^{(t)} \approx A_i$. With this assumption, and our energy function encoding pedestrian velocity preferences (defined in the previous subsection), the state transition of pedestrian i from time t to $t + \Delta t$ is defined by the following dynamical system:

$$\mathbf{p}_i^{(t+\Delta t)} = \mathbf{p}_i^{(t)} + \mathbf{v}_i^{(t+\Delta t)} \Delta t \tag{11}$$

$$\mathbf{v}_i^{(t+\Delta t)} = \underset{\mathbf{v}}{\text{argmin}} E_{\Theta}(\mathbf{v}; s_i^{(t)}, \boldsymbol{s}_{-i}^{(t)}) \; . \tag{12}$$

We use a gradient descent algorithm to solve for the minima of the energy function.

We learn the 8 parameters $\Theta = \{\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \sigma_d, \sigma_w, \beta\}$ required by our energy function from previously annotated data.

In order to make behavioral predictions with our model, we need to deal with the fact that personal properties u_i, $\mathbf{z}_i$ and A_i are unobservable and thus unavailable at prediction time. To deal with this problem, we estimate the personal properties from the past history of states, as described in Section 5.

We learn optimal parameters Θ^* by fitting the energy function to fully observed trajectories in the labeled training data. During training, while predicting the behavior of an individual pedestrian, we fix the states of the other pedestrians to the ground truth. Let us denote the ground truth data by $\tilde{s}_i$. We define the learning problem as computing

$$\Theta^* \in \underset{\Theta}{\text{argmin}} \sum_i \sum_t \left| \tilde{\mathbf{p}}_i^{(t+\Delta t)} - \mathbf{p}_i^{(t+\Delta t)}(s_i^{(t)}, \tilde{\boldsymbol{s}}_{-i}^{(t)}, \Theta) \right|^2 \; . \tag{13}$$

This is a complex nonlinear optimization problem and computing a global optimum is hard. We use a variant of simplex algorithm, with restarts, to solve for a local minima.

We use `eth` and `hotel` sequences from [8] as training data. The dataset includes a total of 750 pedestrians with 15,452 observations of positions under 2.5 frames per second. To estimate the personal properties, we assume the preferred speed of a pedestrian is the average speed over that person's trajectories. The destination is set to be one of 4-5 manually labeled positions outside the scene according to the direction and position at the end of the trajectories.

The social group is also manually labeled. We model scene obstacles as virtual pedestrians with zero-velocity, and manually set these positions along the actual scene obstacles. We sub-sampled at most 12 consecutive tracks (4.8 seconds) every 4 frames (1.6 seconds) for each pedestrian track from each dataset. Then we use these tracks in (13) to learn the parameters.

4. Estimation of personal properties

Our model requires knowledge of the hidden personal properties, preferred speed, u_i, destination, $\mathbf{z}_i$, and social grouping, A_i, for behavioral prediction. As described in more detail in this section, we estimate these variables using the trajectory's history information available at prediction time t.

4.1. Preferred speed

We assume a mean speed of past N_{past} steps as the preferred speed of the person.

A simple, but ineffective alternative is to assume a single global speed for all pedestrians. According to pedestrian speed statistics in [10], a typical person walks around 1.3 m/s. However, this approach ignores individual differences and seems too rough in complex scenes (e.g., sometimes a person slows down to look around, or walks together with kids).

4.2. Destination

The key observation here is that a scene contains only a few types of trajectories. For example, if a scene is a street laying from left to right, we observe persons either passing from left to right or right to left. In this case, it is easy to see that a person walking toward the right side also has his destination in the right side of the scene. This simple assumption might not work if the scene is more complex and has more potential destinations. But looking at certain previous steps in someone's past motion gives us a strong cue as to where his destination is in the scene.

We generalize this observation to the destination prediction problem. Given a past trajectory $r_i^{(t)} = \{s_i^{(t')}\}_{t' \leq t}$, we want to predict a destination $\mathbf{z}_i \in \{Z_1, Z_2, Z_3, ..., Z_K\}$ of the pedestrian i.

We introduce a trajectory feature function $\mathbf{f}_{\text{dest}}(r_i^{(t)})$ and train a K-class classifier C_{dest} to predict the destination:

$$\hat{\mathbf{z}}_i^{(t)} = C_{\text{dest}}(\mathbf{f}_{\text{dest}}(r_i^{(t)})) \ . \tag{14}$$

The feature representation of the trajectory is a concatenation of the normalized histograms of position $\mathbf{p}_i$, speed $|\mathbf{v}_i|$ and direction $\arctan(\mathbf{v}_i)$. In our experiments, position, speed and direction histograms are discretized into 7-by-7, 7 and 9 bins, respectively. All the histograms have equally spaced bins.

We adopt linear support vector machine (SVM) as a classifier for this task. It is generally preferred to use as little trajectory history information as possible to estimate the destination, especially when using behavioral prediction in real time applications. We evaluate the estimation performance with respect to number of past step used to compute features in the next section.

4.3. Social groups

Pedestrians walking in groups tend to behave differently from pedestrian walking alone. Pedestrian in groups tend to walk at the same speed while keeping certain distance between each other. As attempted in [9], we also try to estimate social groups in a scene, but using a simple yet more efficient approach.

The task is to decide whether a pedestrian i and another pedestrian j are in the same group. More precisely, given a pair of past trajectories $(r_i^{(t)}, r_j^{(t)})$, we want to assign a binary label $y_{ij} \in \{+1, -1\}$ that indicates whether they are in the same group $(+1)$ or not (-1). This is a binary classification problem over pairwise trajectories. By defining a feature function $\mathbf{f}_{\text{group}}(r_i^{(t)}, r_j^{(t)})$, we can train a classifier C_{group} from training data:

$$\hat{y}_{ij} = C_{\text{group}}(\mathbf{f}_{\text{group}}(r_i^{(t)}, r_j^{(t)})) \ . \tag{15}$$

The predicted social group is then given by

$$\hat{A}_i = \{j | \hat{y}_{ij} = +1, j \neq i\} \cup \{i\} \ . \tag{16}$$

We use the following quantities as features:

1. normalized histogram of distance $|\mathbf{p}_i - \mathbf{p}_j|$,
2. normalized histogram of absolute difference in speed $||\mathbf{v}_i| - |\mathbf{v}_j||$,
3. normalized histogram of absolute difference in direction $|\arctan(\mathbf{v}_i) - \arctan(\mathbf{v}_j)|$,
4. normalized histogram of absolute difference in velocity direction and relative position $|\arctan(\mathbf{p}_j - \mathbf{p}_i) - \arctan(\mathbf{v}_i)|$, and
5. time-overlap ratio $|T_i^{(t)} \cap T_j^{(t)}| / |T_i^{(t)} \cup T_j^{(t)}|$, where $T_i^{(t)} = \{t' | t' \leq t, s_i^{(t')} \neq \emptyset\}$, i.e., a set of past time steps in which pedestrian i appears.

As with destination estimation, we use an SVM classifier. In the next section, we show the accuracy of prediction as a function of the number of past steps used to produce the feature values.

5. Evaluation

5.1. Datasets

For evaluation, we used the `eth` and `hotel` sequences from [8], and the `zara01`, `zara02` and `stu03` sequences

Table 1. Total number of annotations in datasets

Dataset	eth	hotel	zara01	zara02	stu03
Frames	1448	1168	866	1052	541
Pedestrians	360	390	148	204	434
Groups	243	326	91	140	297
Observations	8908	6544	5024	9537	17583
Destinations	5	4	4	4	4
Obstacles	44	25	34	34	16

Table 2. Destination prediction accuracy

	N_{past}					
Dataset	0	1	2	4	8	∞
eth	75.9	77.2	77.4	77.8	78.2	78.2
hotel	71.2	70.5	70.7	70.3	70.7	71.1
zara01	96.2	96.0	96.0	96.0	95.8	96.7
zara02	82.5	82.5	82.3	82.5	82.4	86.4
stu03	65.6	66.8	66.2	66.8	67.3	66.7

Table 3. Social group prediction precision and recall

		N_{past}					
	Dataset	0	1	2	4	8	∞
Precision	eth	78.6	79.0	79.5	80.3	82.0	83.0
	hotel	86.5	86.1	86.5	86.5	88.5	91.3
	zara01	92.6	90.4	93.6	93.3	94.3	97.2
	zara02	67.4	67.5	67.6	68.0	68.5	88.8
	stu03	70.1	71.6	72.4	74.3	75.9	80.5
Recall	eth	80.2	78.8	79.4	79.3	78.7	77.8
	hotel	91.0	95.3	94.4	95.3	95.9	95.4
	zara01	75.8	78.1	76.6	76.1	76.2	74.9
	zara02	85.0	86.5	87.7	88.3	90.3	92.3
	stu03	73.8	73.0	73.8	74.6	77.0	77.0

from [5]. These sequences have annotated positions. We manually added the annotations of scene obstacles, destinations and social groups. Table 1 summarizes the number of annotations in the different datasets. All the sequences have 25 fps, and annotations are given every 0.4 seconds.

We use all the sequences to evaluate our personal-property estimator. For the experiments on behavioral prediction and tracking, we use `eth` and `hotel` to learn parameters, and `zara01`, `zara02` and `stu03` to evaluate.

5.2. Personal-property estimation

To evaluate the performance of destination and group estimation, we ran a 3-fold cross-validation on prediction accuracy. In this experiment, we do this by subsampling tracks $\{s_i^{(t')}\}_{t-N_{\text{past}}\Delta t \leq t' \leq t}$ for $N_{\text{past}} = \{0, 1, 2, 4, 8, \infty\}$, [3] every 4 frames for each person i. The sampled trajectories are then uniformly split into the 3 sets used for the 3-fold cross-validation. Table 2 shows the average accuracy of destination prediction while Table 3 shows the average precision and recall of social group prediction, both as a function of the number of past steps used to compute trajectory features.

The experimental results in Table 2 suggest that the difficulty of destination estimation depends strongly on the type of scene. Typically, confusion occurs when trajectories having different destinations share a sub-trajectory. In fact, our estimation is worse in the `eth` and `hotel` datasets than in the `zara01` and `zara02`, because the initial part of trajectories in the former datasets look very similar, even if those trajectories later diverge as pedestrians move. The estimation in `stu03` dataset is worst because, in that dataset, many people standing at the same location, which confuses our predictor. Note that in the annotation we automatically assigned the closest destination located outside the scene to pedestrian temporarily standing. Also, we can observe that the number of past steps used to compute trajectory features has almost no influence on prediction accuracy. Rather, it is remarkable that the estimation using only the current state $s_i^{(t)}$ already gives reasonable performance. This indicates that the location and velocity of the pedestrian in the current scene already provides enough information to guess where that person will move in the future.

Table 3 shows that the social group can be estimated with reasonably well, regardless of scene environment. Also, in this case, having more past information does indeed improve estimation performance.

Note that in this experiment we predict a label of a directional link label between two persons but do not consider the symmetric and transitive properties of social relations in groups. Imposing these properties via additional constraints might further improve estimation performance.

5.3. Behavioral prediction

To evaluate the performance of behavioral prediction, we calculated the average displacement of the predicted position of a single pedestrian from the ground truth position. As in parameter learning, in this experiment, we also fix the states of other pedestrians to the ground truth. We evaluated the error in the `zara01`, `zara02` and `stu03` datasets using the parameters learned from the `eth` and `hotel` datasets. Because the destination estimator requires scene specific data, we performed 2-fold cross-validation by splitting each dataset into a first and a second half (corresponding to the initial and last period of time in the video).

Similarly to parameter learning, we subsampled at most 12 consecutive tracks (4.8 seconds) every 4 frames (1.6 seconds) for each pedestrian track, and computed prediction

[3] By $N_{\text{past}} = \infty$ we mean all past steps available.

Table 4. Error in behavioral prediction (m)

	Dataset		
Method	zara01	zara02	stu03
LIN	0.442	0.396	0.556
LTA	0.372	0.376	0.536
LTA+D	0.391	0.378	0.546
ATTR	0.346	0.387	0.532
ATTR+D	0.349	0.373	0.509
ATTR+G	0.347	0.386	0.516
ATTR+DG	0.352	0.367	0.495

error using the average of the objective function given in (13). Tracks in the training set are then used to build personal property estimators. We allow estimators to use at most 5 past steps of information to compute features.

In this evaluation, we compared the constant speed model (LIN); the collision avoidance model of [8] with ground truth destination (LTA) and with predicted destination (LTA+D); and our model with ground truth (ATTR), predicted destination (ATTR+D), predicted social groups (ATTR+G) and predicted destination and social groups combined (ATTR+DG). The preferred speed is set to ground truth in all cases. Table 4 summarizes the average displacement (in meters) from ground truth at each prediction. The result is the average of a 2-fold cross-validation.

Both LTA and our model perform better than LIN in all cases, with or without ground truth. We can also see that using predicted destination and social groups does not degrade the error significantly, and in fact, their combination with our behavioral model produces better results in the `zara02` and the `students03` datasets. This may seem to contradict our intuition, because the model may be using incorrectly predicted destinations. However, those datasets have many crowded scenes in which often pedestrians stop walking to chat or look around. In that case, predicting a destination outside the scene apparently is unreasonable. We believe our predictions dynamically provided more reasonable destinations for such tricky pedestrians and thus better describe the actual situation.

Figure 1 shows an example of the prediction over 12 steps in the `zara01` dataset. A subject is moving towards the rightside of the scene with another person, and is about to pass by another group moving in the opposite direction. The LIN model loses its track from the ground truth. Both LTA and our model track the ground-truth pedestrian path more closely. However, LTA predicts a straight path towards the goal while our model also predicts fluctuations as a consequence of the social interaction.

5.4. Tracking

We evaluated the effect of our behavioral model in a tracking application. Having in mind a video surveillance scenario using a low frame rate webcam, we compare the number of successful tracks achieved by the different models under an observation frequency of every 0.4 seconds (2.5 fps) and 1.6 seconds (0.625 fps), in the `zara01`, `zara02` and `students03` datasets, keeping behavioral prediction running every 0.4 seconds in both cases.

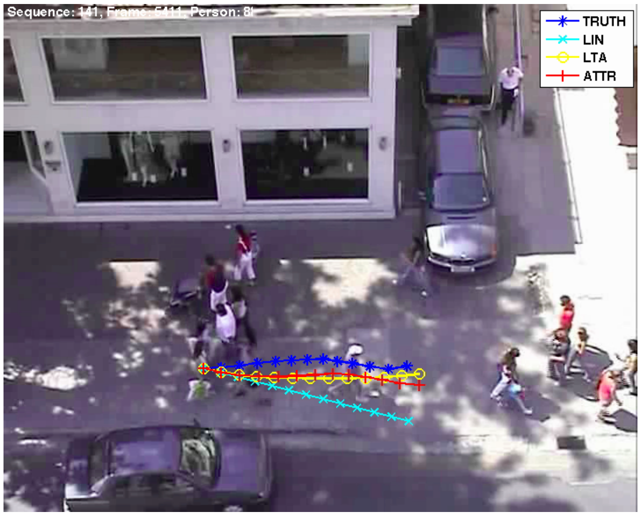

Figure 1. Example of behavioral prediction

To illustrate the positive effect of our behavioral prior in this setting, we use a simple pixel correlation for the appearance model. Our tracker is a simple combination of appearance and behavioral model:

$$P(\mathbf{p}) = P_{\text{appearance}}(\mathbf{p}) \cdot P_{\text{behavior}}(\mathbf{p})$$
$$\propto \exp\left(-\frac{(1-\text{NCC}(\mathbf{p}))^2}{2\sigma_a^2}\right) \exp\left(-\frac{|\mathbf{p}-\mathbf{p}_i^{(t)}|^2}{2\sigma_b^2}\right) .$$

where NCC is the normalized cross correlation of pixels, $\mathbf{p}_i^{(t)}$ is the prediction from the behavioral model, and σ_a and σ_b are the variance parameter for each model. The predicted position at time t is thus given by

$$\hat{\mathbf{p}}_i^{(t)} = \text{argmax}_{\mathbf{p}}\ P(\mathbf{p}).$$

Under the less frequent image observation, we treat the behavioral prediction as the combined prediction when we do not have an appearance term.

The video sequences in the datasets have relatively noisy background. We first apply background subtraction as an image preprocessing step before we compute pixel correlations. We use a running average as background of the scene, and regarded a region having small absolute difference from the model as a background. We start to accumulate background 5 steps before we start tracking in each sequence.

We experiment with tracking in a subsequence of videos. As in the case of the previous section, we split the dataset into a first half and a second half, train a personal-property estimator in one fold, and test in the other. Since our social group estimator is compatible across datasets, in this experiment we use a single group estimator trained using all three datasets. We start tracking every 16 frames and keep them running for at most 24 subsequent frames as long as the ground truth data exist for the scene. The experimental data contains 55, 64 and 34 subsequences for the `zara01`, `zara02` and `students03` datasets, respectively, in total for the 2-fold experiments. The tracker starts from a ground-truth state, with at most 5 past steps available for personal-property prediction and background subtraction. Once the tracker starts, no future or ground truth information is available.

We compare the tracking performance between a linear model with full appearance observation under 2.5 fps (LIN+FULL), with less frequent appearance observation under 0.625 fps (LIN+LESS), LTA model with full or less frequent appearance observation (LTA+FULL, LTA+LESS), our model with full or less frequent appearance observation (ATTR+FULL or ATTR+LESS, respectively), and for reference, a simple correlation tracker without any behavioral prior under full image observation.

In this experiment, we use predicted personal properties for LTA and our model. To comply with the tracking method in [8], the LTA model uses nearest-neighbor decisions to predict destination, using current direction of velocity and direction to a set of destinations. Our model uses the prediction method of section 5.

The evaluation consists of calculating how many trackers stay within 0.5 meter from the ground-truth annotation of the same person at the $N = \{8, 16, 24\}$-th step since the initial state. A track that is more than 0.5 meter away from its corresponding track in the ground-truth data is regarded as *lost*, while a track that is within 0.5 meter from ground-truth but whose closest track in the ground-truth is of different person is considered *ID switch*.

Figure 2 compares the number of successful tracks between tracking methods. The performance of a tracker under full appearance observation (2.5 fps) does not vary among behavioral models, and full appearance observation always results in performance improvement. However, under less frequent observation, our method outperforms in `zara01` and `zara02` dataset. Table 5 summarizes the number of successful, ID-switched, or lost tracks under less frequent appearance observation. The `stu03` result shows that the linear model is the best among others. This is likely the result of irregular type of pedestrians in the dataset: in those scenes, there are many people who walk around unpurposefully and stop to chat. Both LTA and our model assume that a pedestrian is always walking towards the goal and cannot correctly deal with a standing person. This resulted in better performance for the linear model.

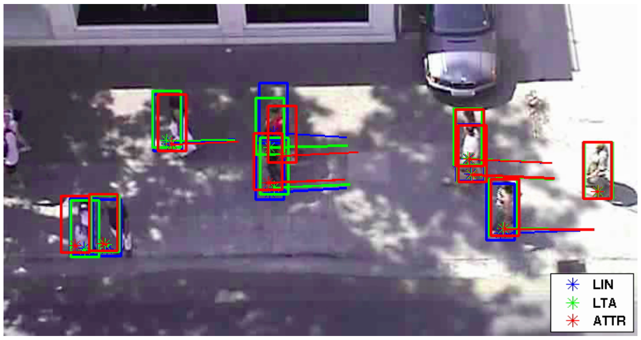

Figure 3. Tracker example. The line indicates the displacement from the previous frame. Under 0.625 fps, it is hard to find a correspondence between frames without prior.

Figure 3 shows an example of a tracker from the `zara01` dataset. Our behavioral model gives stronger preference to keeping the distance between pedestrians in the same social group constant.

6. Conclusion and Future Work

We propose an agent-based formulation of pedestrian behavior and a method to estimate hidden personal properties. Our evaluation of destination and social group estimation, together with that of the behavioral prediction error, suggests that it is possible to get a reasonable estimate of unobservable personal information from purely behavioral and environmental data only. Our tracking experiment shows that, for usual scenes where pedestrians do not exhibit sudden irregular motions, our behavioral model further improves performance over simpler behavioral models under low frame rates.

It would be interesting to extend our behavioral model by using an explicit model of pedestrian behavior that accounts for more that just a walking state. Also, in our future work, we will take into account the interaction between pedestrian behavior and scene events or objects.

References

[1] G. Antonini, S. Martinez, M. Bierlaire, and J. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69:159–180, 2006. 10.1007/s11263-005-4797-0. 1346

[2] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1282 –1289, 2009. 1346

[3] W. Ge, R. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *Applications of Computer Vision (WACV), 2009 Workshop on*, pages 1–8. IEEE, 2010. 1346

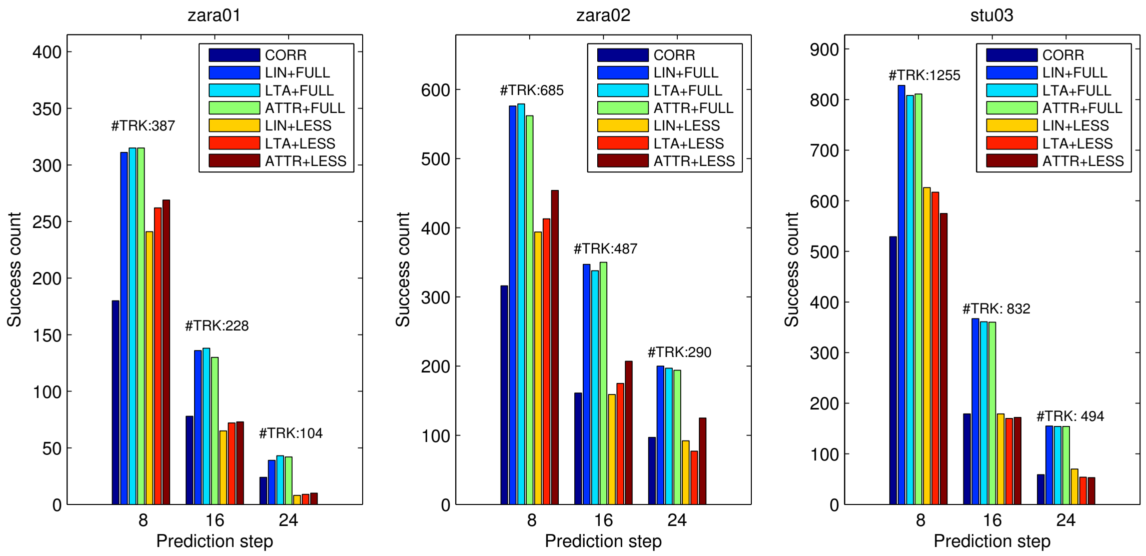

Figure 2. Number of successful tracks over prediction steps that stayed within 0.5 meter from the truth. Under the full appearance observation, the performance does not change among behavioral models. However, appropriate behavioral prior helps improving tracking performance under limited image observation in `zara01` and `zara02`.

Table 5. Tracking result

		zara01			zara02			stu03		
Track	Method	N=8	N=16	N=24	N=8	N=16	N=24	N=8	N=16	N=24
Success	LIN+LESS	241	65	8	394	159	92	**626**	**179**	**70**
	LTA+LESS	262	72	9	413	175	77	617	170	54
	ATTR+LESS	**269**	**73**	**10**	**454**	**207**	**125**	575	172	53
ID switch	LIN+LESS	18	19	**9**	37	30	15	**98**	92	**50**
	LTA+LESS	**13**	20	11	38	**26**	13	**98**	91	58
	ATTR+LESS	18	**17**	**9**	**36**	40	**12**	99	**87**	56
Lost	LIN+LESS	128	144	87	254	298	183	**531**	**561**	**374**
	LTA+LESS	112	**136**	**84**	234	286	200	540	571	382
	ATTR+LESS	**100**	138	85	**195**	**240**	**153**	581	573	385

[4] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Phys. Rev. E*, 51(5):4282–4286, May 1995. 1346

[5] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. *EUROGRAPHICS*, 2007. 1346, 1349

[6] M. Luber, J. Stork, G. Tipaldi, and K. Arras. People tracking with human motion predictions from social forces. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 464–469. IEEE, 2010. 1346

[7] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, 2009. 1346

[8] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. *International Conference on Computer Vision (ICCV)*, 2009. 1346, 1347, 1348, 1350, 1351

[9] S. Pellegrini, A. Ess, and L. van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. *ECCV 2010*, 2010. 1346, 1348

[10] T. Robin, G. Antonini, M. Bierlaire, and J. Cruz. Specification, estimation and validation of a pedestrian walking behavior model. *Transportation Research Part B: Methodological*, 43(1):36 – 56, 2009. 1346, 1348

[11] P. Scovanner and M. Tappen. Learning pedestrian dynamics from the real world. *International Conference on Computer Vision (ICCV)*, pages 381–388, 2009. 1346

[12] H. Timmermans. *Pedestrian Behavior: Data Collection and Applications*. Emerald Group Publishing Limited, 2009. 1346

[13] B. Ziebart, N. Ratliff, G. Gallagher, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa. Planning-based prediction for pedestrians. *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009. 1346