You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking

S. Pellegrini[1], A. Ess[1], K. Schindler[1,2], L. van Gool[1,3]

[1] Computer Vision Laboratory, ETH Zurich, Switzerland
[2] Computer Science Dept., TU Darmstadt, Germany
[3] ESAT/PSI-VISICS IBBT, KU Leuven, Belgium
{stefpell|aess|konrads|vangool}@vision.ee.ethz.ch

Abstract

Object tracking typically relies on a dynamic model to predict the object's location from its past trajectory. In crowded scenarios a strong dynamic model is particularly important, because more accurate predictions allow for smaller search regions, which greatly simplifies data association. Traditional dynamic models predict the location for each target solely based on its own history, without taking into account the remaining scene objects. Collisions are resolved only when they happen. Such an approach ignores important aspects of human behavior: people are driven by their future destination, take into account their environment, anticipate collisions, and adjust their trajectories at an early stage in order to avoid them. In this work, we introduce a model of dynamic social behavior, inspired by models developed for crowd simulation. The model is trained with videos recorded from birds-eye view at busy locations, and applied as a motion model for multi-people tracking from a vehicle-mounted camera. Experiments on real sequences show that accounting for social interactions and scene knowledge improves tracking performance, especially during occlusions.

1. Introduction

Object tracking has seen considerable progress in recent years, with current systems able to handle long and challenging sequences automatically with high precision. The progress is mostly due to improved object models—either generic appearance models or detectors for specific kinds of objects—or better optimization strategies. One aspect that was hardly explored so far however is the dynamic model, another key component of every tracking approach. Typically, a standard first-order model is used, which does not account for the real complexity of human behavior.

In particular, physical exclusion in space is often modeled only indirectly, by allowing at most one detection to be assigned to a trajectory, while at the same time making sure that detections are sufficiently far from each other. In practice this amounts to non-maximum suppression in 2D image space. In situations where full occlusions are common (*e.g.* in street scenes seen by a street-level observer), such an image-based approach fails to adequately differentiate collisions from occlusions.

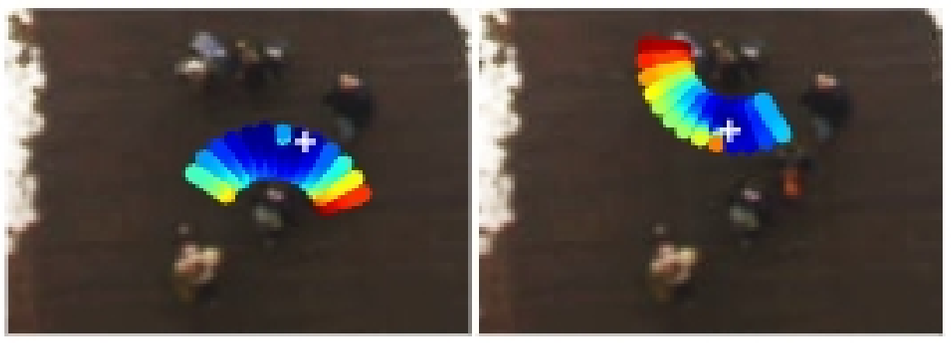

Figure 1. While walking among other people, several factors influence short-term path planning. Smoothness of motion, intended destination, and interactions with others limit one's choice of direction and speed. In the example (same scene, two pedestrians' perspectives), blue indicates good choices for velocity, red signals "no-go"s. The white cross shows the actually chosen velocity. We propose a dynamic model that takes these factors into account.

We believe that one main problem in this context is the dynamic model, typically a first- or second-order approximation applied *independently* to each subject, *e.g.* using an Extended Kalman Filter (EKF). Inspired by work on crowd simulation, we propose a more elaborate dynamic model, which takes into account the social interactions between objects (here, pedestrians) as well as their orientation towards a destination (usually outside the field of view). The fact that people proactively anticipate future states of their environment during path planning, rather than only react to others once a collision is imminent, has largely been ignored in the literature. This goes to the extent that standard motion models do not even take into account the elementary fact that people have a destination, and hence steer back to their desired direction after deviating around an obstacle.

The proposed model, termed *Linear Trajectory Avoidance* (LTA), is designed for walking people with short-term prediction in mind. Due to the complexity of human motion patterns, longer prediction horizons become unreliable; *very* short ones do not require sophisticated models, since displacements are so small that linear extrapolation is sufficient. Hence, the effect of LTA is best seen in busy scenar-

ios with frequent short-term occlusions, or when framerate is low and the data association procedure is less reliable.

The model (Sec. 3) operates in physical world coordinates and can be applied to any tracker which operates in a metric frame. We show how the model parameters can be learned from birds-eye view data(Sec. 4), and apply it both in a simple patch-based tracker operating on oblique views, and in a detection-based tracker operating on footage from a moving camera (Sec. 5).

2. Related Work

Multi-target tracking. In recent years, object tracking has been successfully extended to scenarios with multiple objects [12, 16, 19]. Modern systems can track through long and challenging sequences with high precision. To this end, researchers have focused on improving the appearance model [10, 5], the object detector [2, 7, 9, 22], and/or the optimization strategy [14, 16, 23]. Others have developed approaches specifically for crowded scenes [1, 6, 24].

The dynamics and interaction between targets is much less explored. Several models include the requirement that the tracked objects should not collide in any frame. The condition is met by assigning every object detection to at most one tracked object [12, 19, 22]. Note that the unique assignment alone does not solve the problem for finite object size and finite framerate: detections are not guaranteed to be far enough apart to prevent collisions—one has to rely on non-maximum suppression in image space. Furthermore, there are valid assignments which give rise to crossing paths with a collision between adjacent frames.

In their "space-time event-cone tracking", Leibe *et al.* [16] explicitly model physical exclusion between subjects in world coordinates, however, this is restricted to the selection of the best trajectory hypotheses only—the important step of creating these hypotheses is done independently and does not cater for interactions.

Besides interactions, one important factor in our model is the desired direction of a subject by the way of goal points. Such points have been used to influence tracking [1, 12, 14]. We directly include target points in our optimization.

Social behavior models. Modeling the behavior of pedestrians has been an important area of research mainly in evacuation dynamics and traffic analysis. Pedestrian behaviors have been studied from a crowd perspective, with *macroscopic* models for pedestrian density and velocity. On the other end of the spectrum, *microscopic* models deal with individual pedestrians. One example for the latter is the *social force* model [11], where pedestrians react to energy potentials caused by other pedestrians and static obstacles through a repulsive force, while trying to keep a desired speed and motion direction. Another branch of microscopic models assumes *agents* that interact autonomously through a basic form of intelligence represented by a rule set [15, 20]. In yet another branch, cellular automata are used, which discretize the space and select the next desired direction from a preference matrix, *e.g.* [21].

All these models have been designed and used for simulation purposes. This is also the case for the example-based model of Lerner *et al.* [17], although in this work the simulation is used for synthesizing computer graphics videos.

We are only aware of three works, which use a pedestrian model in computer vision applications. Ali and Shah [1] use the cellular automaton model atop a set of scene-specific "floor fields" to make tracking in extremely crowded situations tractable. In contrast, we model single pedestrians in world coordinates, which decouples the approach from the camera setup. Antonini *et al.* [3] propose a variant of the Discrete Choice Model to build a probability distribution over pedestrian positions in the next time step, assuming that all subjects perform a global optimization for the next step. Very recently, Mehran *et al.* [18] use the social force model to detect abnormal behavior in crowded scenes.

Our LTA model shares some characteristics with the *social force* model [11], but differs in two crucial ways: first, rather than modeling the pedestrians at their current location as energy potentials, we predict their *expected point of closest approach*, and use that point as the driving force for decisions. Second, when simulating a subject, we make it move in the optimal direction instead of just applying a gradient-dependent force. Hence, in LTA pedestrians exhibit decisive behavior and choose their path such as to minimize collisions, rather than just being reactive particles.

3. Modeling Social Behavior

Given a current configuration $\mathcal{S} = \{s_i\}$ of subjects $(i = 1 \dots n)$, our model estimates the velocity of each s_i in the next time step, based on current positions and velocities for all the subjects. Specifically, we model a subject as $s_i = (\mathbf{p}_i^t, \mathbf{v}_i^t)$, where $\mathbf{p}_i^t$ denotes its 2D position on the ground plane and $\mathbf{v}_i^t$ its velocity vector at time t. For brevity's sake, we define the current time step as $t = 0$ and drop the corresponding superscript, *e.g.* $\mathbf{p}_i = \mathbf{p}_i^0$. In the following, we will first concentrate on the basic case of two subjects before generalizing to an arbitrary number.

We assume a first-order model jointly for all pedestrians in the scene: every pedestrian knows the current positions and velocities of all subjects. It is thus reasonable to think that each pedestrian will predict the movement of the other pedestrians following a constant velocity model. Therefore, if subject s_i proceeds with the velocity $\bar{\mathbf{v}}_i$, it expects to have the squared distance $d_{ij}^2(t)$ from s_j at time t:

$$d_{ij}^2(t, \bar{\mathbf{v}}_i) = \|\mathbf{p}_i + t\bar{\mathbf{v}}_i - \mathbf{p}_j - t\mathbf{v}_j\|^2 \quad , \qquad (1)$$

where we have made explicit the dependence of the d_{ij} to $\bar{\mathbf{v}}_i$ to highlight that we are taking the perspective of s_i

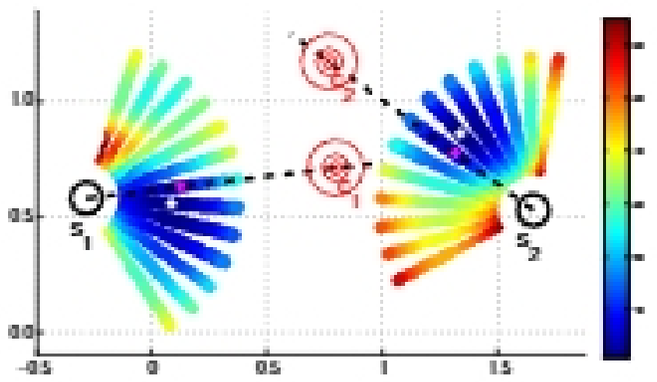

Figure 2. Two subjects, with their current directions (black) and velocities (magenta). s_1 feels the repulsion from s_2's expected point of closest approach c_2, and vice versa. Colors denote energies for different velocities, white dots mark the respective minima. Note how s_2 accelerates and turns right in order to avoid s_1, while s_2 slows down and turns to his right.

(w.l.o.g.). Defining $\mathbf{k}_{ij}^t = \mathbf{p}_i^t - \mathbf{p}_j^t$ and $\mathbf{q}_{ij}^t = \bar{\mathbf{v}}_i - \mathbf{v}_j^t$ we can rewrite Eq. 1 as

$$d_{ij}^2(t, \bar{\mathbf{v}}_i) = \|\mathbf{k} + t\mathbf{q}\|^2 \quad . \tag{2}$$

We assume that pedestrians try to steer clear of collisions. As s_i has an estimate for s_j's velocity from the last time step, it will adapt its own velocity $\bar{\mathbf{v}}_i$ such that the minimum distance d_{ij}^{*2} from s_j is greater than a certain value that s_i considers *comfortable*. The minimum distance occurs at the time of closest approach t^*, where

$$t^* = \underset{t>0}{\operatorname{argmin}}\, d_{ij}^2(t, \bar{\mathbf{v}}_i) \quad , \tag{3}$$

and we constrain the search to future time steps. Relaxing this constraint for a moment, the time at which the distance is minimized is found by setting the derivative of d_{ij}^2 with respect to t to zero,

$$\frac{\partial d_{ij}^2(t, \bar{\mathbf{v}}_i)}{\partial t} = 2(\mathbf{k} + t\mathbf{q})\mathbf{q}^\top = 0 \quad \rightarrow \quad t^* = -\frac{\mathbf{k} \cdot \mathbf{q}}{\|\mathbf{q}\|^2} \, . \tag{4}$$

In Eq. 4, the distance d_{ij}^2 decreases for $t < t^*$ and increases for $t > t^*$. We can therefore reintroduce the constraint, saying that if t^* is smaller than zero, then the minimum of d_{ij}^2 for $t \geq 0$ will be at $t = 0$. Substituting Eq. 4 into Eq. 2 then yields the minimum distance

$$d_{ij}^{*2}(\bar{\mathbf{v}}_i) = \|\mathbf{k} - \frac{\mathbf{k} \cdot \mathbf{q}}{\|\mathbf{q}\|^2}\mathbf{q}^\top\|^2 \quad . \tag{5}$$

Note that Eq. 5 does not depend on time anymore. In order to make sure that s_i avoids s_j, one could set Eq. 5 equal to some preferred distance. However, this does not extend well to the case of multiple pedestrians. We therefore propose to build an energy functional for the interaction between s_i and s_j as a function of d_{ij}^{*2},

$$E_{ij}(\bar{\mathbf{v}}_i) = c^{-\frac{d_{ij}^{*2}(\bar{\mathbf{v}}_i)}{2\sigma_d^2}} \quad , \tag{6}$$

where σ_d controls the distance to the subject to be avoided. E_{ij} is maximal when the linear trajectories would lead to a collision, and is minimal as d_{ij}^{*2} goes to infinity.

Based on Eq. 6, the influence of multiple subjects can now be modeled as a weighted sum, where each subject s_r $(r \neq i)$ gets assigned a weight $w_r(i)$ depending on its current distance and angular displacement ϕ from s_i. We set

$$w_r(i) = w_r^d(i) w_r^\phi(i) \tag{7}$$

$$w_r^d(i) = c^{-\frac{\|\mathbf{k}_{ir}\|^2}{2\sigma_w^2}} \tag{8}$$

$$w_r^\phi(i) = \left((1 + \cos(\phi))/2\right)^\beta \quad . \tag{9}$$

σ_w defines the radius of influence of other objects, β controls the "peakiness" of the weighting function used for the field-of-view. The overall interaction energy for subject s_i, $I_i(\bar{\mathbf{v}}_i)$, is then given by

$$I_i(\bar{\mathbf{v}}_i) = \sum_{r/i} w_r(i) E_{ir}(\bar{\mathbf{v}}_i) \quad . \tag{10}$$

These interactions alone, however, do not bound the minimization appropriately because scene knowledge is ignored. Like in other works [1, 13], we assume that each pedestrian walks towards a destination $\mathbf{z}_i$, and in doing so tries to maintain a desired speed u_i. These two components can be represented by two further energy potentials,

$$S_i(\bar{\mathbf{v}}_i) = (u_i - \|\bar{\mathbf{v}}_i\|)^2 \tag{11}$$

$$D_i(\bar{\mathbf{v}}_i) = -\frac{(\mathbf{z}_i - \mathbf{p}_i) \cdot \bar{\mathbf{v}}_i}{\|\mathbf{z}_i - \mathbf{p}_i\| \cdot \|\bar{\mathbf{v}}_i\|} \, . \tag{12}$$

The overall energy for subject s_i can hence be written

$$E_i(\bar{\mathbf{v}}_i) = I_i(\bar{\mathbf{v}}_i) + \lambda_1 S_i(\bar{\mathbf{v}}_i) + \lambda_2 D_i(\bar{\mathbf{v}}_i) \quad , \tag{13}$$

with λ_1 and λ_2 controlling the influence of the two regularizers. See Fig. 1 and Fig. 2 for a visualization of the obtained energies. Minimizing this distance with respect to the velocity $\bar{\mathbf{v}}_i$ cannot be done in a closed form. In our experiments we employ gradient descent with line search.

Given the situation of a pedestrian facing a group of people, an interesting outcome emerges from Eq. 10 and Eq. 13. Fig. 3 shows the energy that a subject s_1 sees when trying to avoid two oncoming pedestrians, s_2 and s_3. Each column of the figure describes the energy for a different direction of the velocity vector (keeping the speed fixed), while each row indicates different distance between s_2 and s_3. One can see that as a consequence of the Gaussian shape, a local minimum in the middle exists only when the gap between the two oncoming subjects is sufficiently large. As the gap narrows, the two people form a local maximum that s_1 will try to avoid.

The minimization of the energy functional allows for the calculation of the next *desired* velocity $\bar{\mathbf{v}}_i^*$. However, due to inertial constraints, the subject has to undertake a transition from the current velocity to the desired one. This is modeled through a simple filtering approach. The subject's position is updated according to

$$\mathbf{p}_i^{t_N} = \mathbf{p}_i + (\alpha_N \mathbf{v}_i + (1 - \alpha_N)\bar{\mathbf{v}}_i^*)\, t_N \quad , \tag{14}$$

where the prediction interval N is made explicit to allow for the adaptation to different frame rates, and α is a mixture