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.

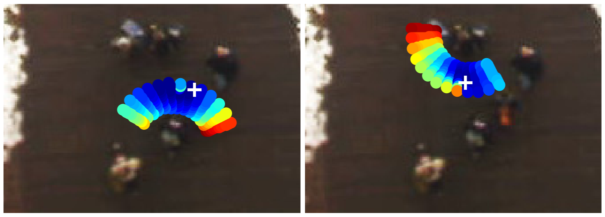

Figure 1. While walking among other people, several factors influence short-term path planning. Smoothness of motion, intended destination, and interactions with others limit one's choice of direction and speed. In the example (same scene, two pedestrians' perspectives), blue indicates good choices for 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-

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

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 $(i = 1 \ldots n)$, 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 $\tilde{\mathbf{v}}_i$, it expects to have the squared distance $d_{ij}^2(t)$ from s_j at time t:

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

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

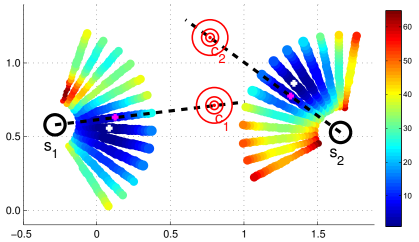

Figure 2. Two subjects, with their current directions (black) and velocities (magenta). s_1 feels the repulsion from s_2's expected point of closest approach c_2, and vice versa. Colors denote energies for different velocities, white dots mark the respective minima. Note how s_2 accelerates and turns 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 $\mathbf{q}_{ij}^t = \tilde{\mathbf{v}}_i - \mathbf{v}_j^t$ we can rewrite Eq. 1 as

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

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 $\tilde{\mathbf{v}}_i$ 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

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

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,

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

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

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

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},

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

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

$$\begin{aligned} w_r(i) &= w_r^d(i) w_r^\phi(i) && (7)\\ w_r^d(i) &= e^{-\frac{||\mathbf{k}_{ir}||^2}{2\sigma_w^2}} && (8)\\ w_r^\phi(i) &= \left((1+\cos(\phi))/2\right)^\beta \quad . && (9)\end{aligned}$$

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, $I_i(\tilde{\mathbf{v}}_i)$, is then given by

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

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,

$$\begin{aligned} S_i(\tilde{\mathbf{v}}_i) &= (u_i - ||\tilde{\mathbf{v}}_i||)^2 && (11)\\ D_i(\tilde{\mathbf{v}}_i) &= -\frac{(\mathbf{z}_i - \mathbf{p}_i)\cdot\tilde{\mathbf{v}}_i}{||\mathbf{z}_i - \mathbf{p}_i||\cdot||\tilde{\mathbf{v}}_i||} \quad . && (12)\end{aligned}$$

The overall energy for subject s_i can hence be written

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

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 $\tilde{\mathbf{v}}_i$ 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 $\tilde{\mathbf{v}}_i^*$. 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

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

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

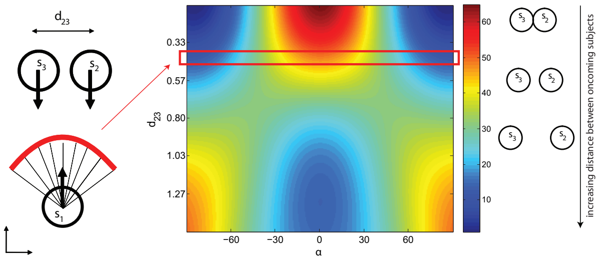

Figure 3. Energy seen by subject s_1 when making a choice of changing its heading (horizontal axis) as it approaches two subjects moving in opposite direction. The wider the gap between s_2 and s_3 (vertical axis), the easier it is to pass between them (bottom of graph, minimum in middle) instead of steering around the pair (top, minima on the side).

coefficient. Naturally, as N grows the prediction becomes more linear. We keep the time interval N at the frame rate of the respective sequence and recompute the desired velocity at each time step.

3.1. Static Obstacles

So far, we only took dynamic obstacles in the form of pedestrians into account. In most common scenes however, people will also try to avoid static obstacles. Following other authors [13] we model such obstacles as subjects with zero velocity. The obstacle's position is approximated at every time step by the point closest to the pedestrian [1, 13]. While being a coarse approximation, this works well except for highly non-convex obstacles.

3.2. Application of the Model

Given the current configuration of dynamic and static obstacles at time t, we infer the optimal velocity at time $t+1$ for each subject in turn by minimizing Eq. 13 and then applying Eq. 14. Once these velocities have been identified for each subject, they are updated in parallel. In the case of tracking, if an observation is available, it is merged with the simulation's estimate at this point. Note that we do *not* iterate the simulation in the current time step, assuming that pedestrians base their immediate path planning only on the past. Also, oscillations *over time* can still occur when people walk towards each other—a well-known situation from everyday life.

4. Training

The model as defined in the previous section has six free parameters, which need to be learned from training sequences: the standard deviations defining the comfortable distance σ_d and the radius of interest σ_w, the "peakiness" β of the subject's field of view, the importance weights λ_1 and λ_2 of the desired speed and velocity, and the update rate α. We fix the prediction time step to 0.4 seconds, which is a reasonable horizon for the model to operate.

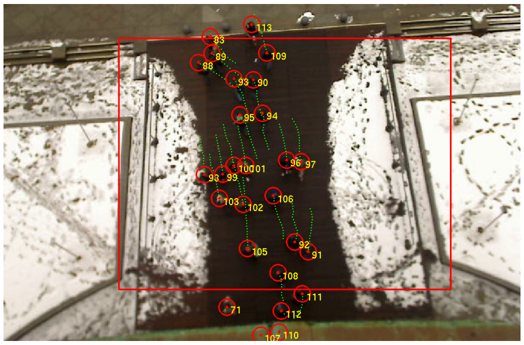

Figure 4. Sample frame from one of the training sequences.

σ_d	σ_W	λ_1	λ_2	β	α
0.361	2.088	2.33	2.073	1.462	0.730

Table 1. Model parameters obtained from training sequences.

To train our model, we have recorded two data sets from birds-eye view and annotated them manually. This gave a total of 650 tracks over 25 minutes.[1] A sample image including annotation can be seen in Fig. 4.

In both scenes, goal points were labeled and the desired direction for each subject was set towards the closest goal. For each pedestrian, the desired speed was set to the mode of his speed histogram. The field-of-view was restricted to ± 90 degrees (i.e., $w_r^\phi = 0$ for $|\phi| > \frac{\pi}{2}$). People standing or strolling aimlessly were ignored.

To find an optimal set of parameters we have experimented with two optimization strategies, namely gradient descent starting from multiple random initializations, and a variant of genetic algorithms (GA). We found that among the returned local optima of the parameters vector, several performed equally well. For the following experiments, we always use the local optimum with the lowest error (which resulted from the GA optimization).

In one iteration round, each subject is simulated in turn, holding the others fixed at the ground truth. The simulation is started every 1.2 seconds along the subject's path, and continues for 4.8 seconds, similar to [13]. The sum of squared errors (distances to ground truth) over all simulations in the round is minimized.

We obtained the parameters given in Tab. 1. At first glance, $\sigma_d = 0.36$ looks reasonable, stating that people will not feel uncomfortable with a person more than ≈ 1 meter away; $\sigma_w = 2.1$ means that people further away than ≈ 6 meters do not influence path planning; β suggests a relevant peak of attention in the center of the field of view. Note that the restricted field-of-view and the value of σ_w imply that pedestrians are actually only aware of a limited portion of the scene.

[1] Data and videos available at www.vision.ee.ethz.ch/~stefpell/lta

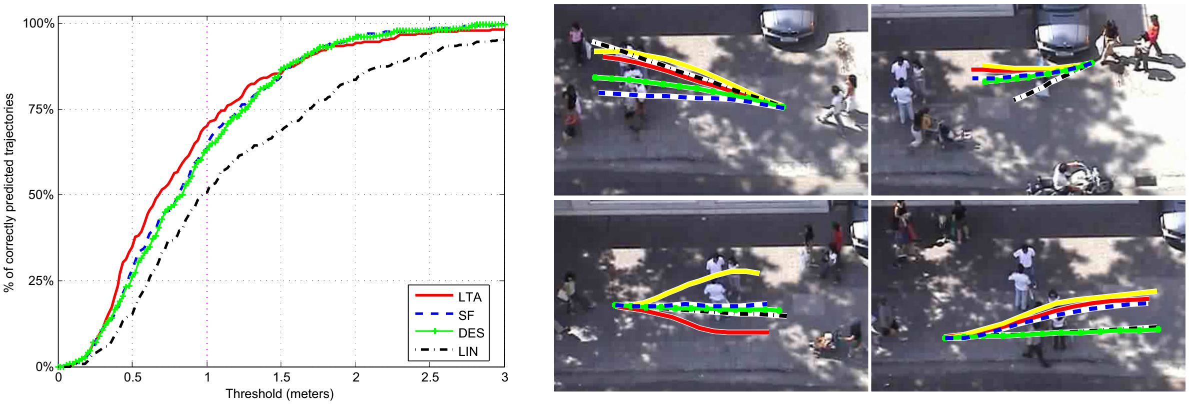

Figure 5. *Left:* performance of the LTA model (solid red) against a trained model that uses destinations but no interactions (crossed green), the social force model [13] (dashed blue) and simple linear extrapolation (dashdot black). *Right:* Example extrapolations: the model smoothly avoids the standing crowd (*top left*, yellow=groundtruth), sometimes however suggests meaningful, but wrong paths (*bottom left*). Using only goal-directed prediction is effective in some cases (*top right*) but in general better prediction is obtained by taking into account interaction among pedestrians (*bottom right*).

5. Results

To experimentally evaluate the trained dynamic model, we test it in three different settings. First, we measure its mere quality as a predictor, which is *e.g.* of interest for path planning in robotics. Then, we apply it inside two tracking methods, a simple patch-based tracker, as well as a state-of-the-art multi-person tracking system.

5.1. Prediction

To test the prediction performance of our model, we use annotated data provided by the authors of [17]. The video shows part of a shopping street from an oblique view. We evaluate on a subsequence of about 3 minutes @ 2.5 FPS containing 86 trajectories annotated with splines. With the same simulation setting used during training (see Sec. 4) this yields $\approx$ 300 simulations. A homography from image to ground plane was estimated from four manually clicked points on the footpath to transfer image to world coordinates. As destinations we chose two points far outside the left and right image borders, which holds for most subjects.

We compare our model with a simple baseline ("LIN"), that merely extrapolates using the previous velocity, and with a re-implementation of the *social force* model ("SF") with elliptical potentials [13]. Parameters for the latter are learned using the procedure discussed in Sec. 4. For our LTA model, we explore two possible parameter sets: the first one was trained without interaction term, adding only the drive towards a destination ("DEST"), whereas the other one ("LTA") also caters for interaction among subjects.

As error measure, the average Euclidean distance between predictions and ground truth is measured in each simulation step. The experiments show an improvement of 6 % in prediction error for the LTA model compared to SF and DEST, and of 24% compared to the LIN model. A closer look at the distribution of the errors sheds more light on the differences between models. For this purpose, we define a trajectory as *correctly predicted* when for each timestep of its simulation, the distance from prediction to ground truth lies within a threshold T. The curve in Fig. 5(left) shows the result of this analysis, plotting the percentage of the correctly predicted trajectories over varying T. At a threshold of 1 meter, $\approx$50% of the trajectories are already correctly predicted using linear extrapolation (LIN). Adding goal-direction (DES) increases the correctly predicted trajectories to $\approx$63%. The SF model performs only slightly better than the DES model. Another $\approx$7% boost is achieved using LTA, reaching a total of $\approx$70%.

There are two issues to note here. Firstly, the scene is only moderately crowded, and a large part of the trajectories are almost straight. For these, all models give satisfactory results, which washes out the average difference. Secondly, the error distribution of LTA has a light but long tail with a small number of very large errors. These happen when the model in its present deterministic form avoids other pedestrians by walking around the wrong side, see Fig. 5. Although from a tracking perspective, bumping into an obstacle is a no less severe failure than passing it on the wrong side, the latter adds twice as large errors and thereby distorts the comparison. A stochastic variant of our model could help here.

5.2. Patch-based Tracking

To highlight the effect of the dynamic model and compare it to the LIN model, we have implemented a simple patch-based tracker, using the normalized cross-correlation (NCC) as similarity measure. In the first frame a rectangular patch is manually initialized at each person's location $\mathbf{p}_i^0$ as appearance model, and the speed of all targets is initialized to $\|\mathbf{v}_i\| = 0$. At each new time step t, the target location $\mathbf{p}_i^t$ is predicted with the dynamic model, and a Gaussian centered at the prediction gives the location prior $P_{pred}(\mathbf{p}) = \frac{1}{Z} \exp\big(-(\frac{\|\mathbf{p}-\mathbf{p}_i^t\|}{2\sigma_{pred}})^2\big)$. In the surroundings of the predicted location, the squared exponential $P_{data}(\mathbf{p}) = \frac{1}{Y} \exp\big(-(NCC(\mathbf{p}, \mathbf{p}_i^0) - 1)^2\big)$ is employed as data likelihood, and the maximum of the posterior $P_{pred} \cdot P_{data}$ gives the new target location.

This simple tracker was applied to short, interesting subsequences of the footpath sequence (non-overlapping with the ones used above). For the dynamic model, we plug in either the LIN (constant velocity) model or our LTA model, leaving the other parameters unchanged. For the LTA model, the desired direction (standing, left-to-right, or right-to-left) is set for each person according to their last displacement, and the desired speed is set to a constant value for all people.

Tracking was performed at 2.5 FPS, leaving 0.4 seconds between consecutive frames. In this scenario with low framerate, multiple interactions, and low data quality, a strong dynamic prior is important to enable tracking at all. As can be seen in the examples of Fig. 6, the simple constant-velocity model loses track of several targets, when they pass others and have to adjust their speed and direction. The examples also show how the trajectories fail to swing back without a target direction. On the other hand, LTA successfully tracks all people in the two examples.

5.3. Tracking with a Moving Observer

To further demonstrate the versatility of the approach, we apply the LTA model (as learned from birds-eye view) to tracking from a moving observer. We use the tracking-by-detection framework [8], and plug in both the LIN and the LTA models for modeling pedestrian dynamics. Both versions are then evaluated on two (publicly available) sequences from that work.

In a nutshell, the approach projects the output of a pedestrian detector—in our case the HOG framework [7]—to 3D world coordinates with the help of visual odometry and a ground plane assumption. The tracking system then generates a set of trajectory hypotheses based on the object detections and a dynamic model, and prunes that set to a minimal consistent explanation with model selection. This pruning relies on the assumption that all actual trajectories are present in the set of hypotheses, thus requiring correct tracking even when no data is available to immediately correct the motion model, mainly during to occlusions. Here the LTA model comes into play.

To make the method amenable to our problem formulation, we adapt as follows: first, instead of creating all trajectory hypotheses independently, we introduce a trajectory extension step that updates all currently active object hypotheses in parallel, making them fight for available detections similar to the greedy approaches used by [22, 19]. This then allows the second, crucial change: in the extension step, we apply the LTA model for each hypothesis in turn, making them anticipate the other subjects' movements in order to avoid them. Especially during occlusion, this ensures that blind trajectory extrapolation takes into account other subjects, and increases the chance that a subject's trajectory leaves the occlusion at the right position, so that tracking can continue correctly. To also detect static obstacles, we additionally project the depth map from stereo images onto a polar occupancy map.

LTA requires a desired orientation and velocity. Assuming very little scene knowledge, we set the desired orientation parallel to the road, pointing in the respective pedestrian's previous direction. The desired velocity is set to the last measured speed of the hypothesis.

As the tracker builds on a quite reliable set of pedestrian detections, we expect an advantage of the LTA model mainly in case of occlusion. The improvement is therefore bounded by the frequency of occlusion events. Then, LTA's extrapolation which is constrained by other agents should outperform a standard linear model, thus preventing possible data association problems when the occlusion is over.

To quantitatively relate the two approaches with each other, we compare tracking output with annotated ground-truth using the CLEAR evaluation metrics [4], which measure ID switches and the percentage of false negative / false positive bounding boxes. In Tab. 2, we compare the two dynamic models by varying the threshold on the Mahalanobis distance d used in the data association. The reasoning behind this procedure is the intuition that a larger search area could possibly compensate for the disadvantages of a less accurate prediction. When using LTA, the number of ID switches is constantly lower, while the number of misses and false positives stays about the same. While consistent, the automatic evaluation tends to over-estimate the number of ID-switches with increasing number of occlusion events. For $d = 3$, we thus manually re-counted the ID switches for the two sequences. In the first sequence, using LTA yields 31 as opposed to 36 ID switches with LIN. In the second sequence, these figures are 18 (LTA) and 26 (LIN). Here, many people leave the field of view and enter again, which is always flagged as a new ID by the tracker. Leaving out these "unrecoverable" cases, the last comparison gets down to 10 (LTA) vs. 18 (LIN), a 44% improvement.

A few interesting situations from the two sequences are shown in Fig. 7. The first three columns show the sequence including the occlusion event as tracked by LTA, then two plots in birds-eye view contrast the results for LTA with those for LIN. Note the ID switches (red arrows), and the missing track in the third example. This last example is especially interesting, because the person in the very front

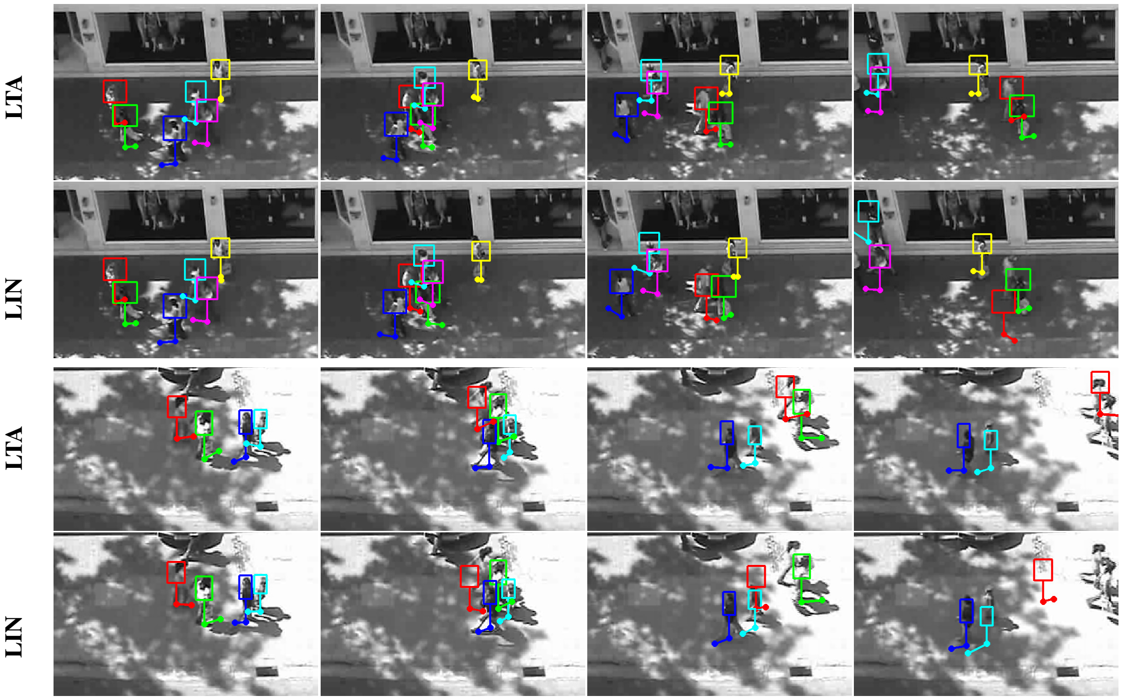

Figure 6. LTA model vs. constant velocity (LIN) model. Selected frames from two tests with the patch-based tracker. *Top:* When using the LTA model, the pedestrian marked in red is constrained by people walking nearby. The LIN model overshoots when he maneuvers around an oncoming person and loses track. *Bottom:* the LIN model for the person marked in red makes a significantly wrong prediction and loses track, whereas the LTA model tries to avoid oncoming people and predicts correctly. Note also how in both examples the persons marked in cyan drift away at the end, because they are not steering towards a target direction.

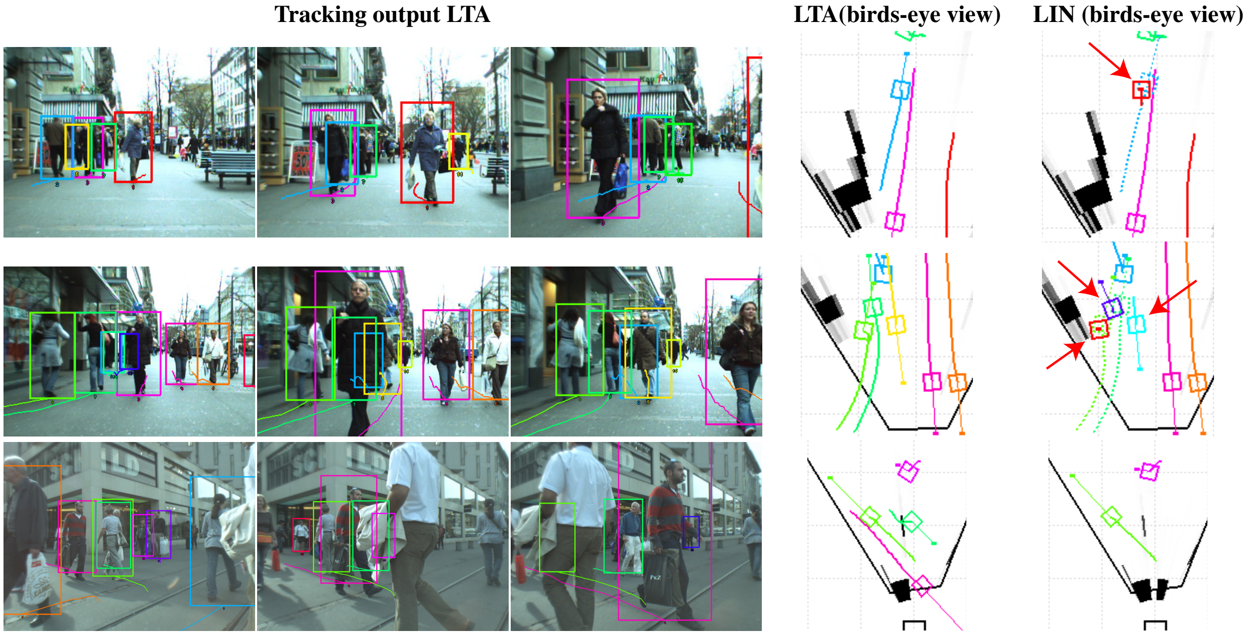

Figure 7. Examples where LTA improves the performance of multi-body tracking. *First three columns:* short sequences with occlusion events, tracking results with LTA. *Column 4:* birds-eye view for the middle frame, using LTA. *Column 5:* birds-eye view for the same frame, using linear model. Black areas are static obstacles, red arrows mark ID switches, dotted lines show the pre-switch trajectories still being extrapolated—these dissappear after ≈5 frames as they fail to find supporting detections. *First row:* the man on the left is successfully recovered from occlusion. *Second row:* constrained by the oncoming person, both ladies and the oncoming man are picked up again. *Third row:* while the man in the front is not detected, he is integrated into the obstacle map, thus constraining the man in the red-black sweater.

	ID switches				misses				false positives			
	1.5	2.0	2.5	3.0	1.5	2.0	2.5	3.0	1.5	2.0	2.5	3.0
	Seq#1											
LIN	55	55	51	48	0.29	0.28	0.28	0.28	0.19	0.19	0.19	0.19
LTA	48	42	45	41	0.28	0.28	0.28	0.28	0.19	0.19	0.19	0.19
	Seq#2											
LIN	35	33	31	31	0.21	0.21	0.21	0.21	0.08	0.08	0.09	0.09
LTA	31	30	26	25	0.21	0.21	0.20	0.20	0.08	0.09	0.08	0.09

Table 2. Comparison of the dynamic models for differing data association thresholds based on the CLEAR evaluation metrics.

is only detected as a static obstacle. Nevertheless it influences the man in the striped sweater, who successfully steers around it, whereas LIN looses track.

6. Conclusion

We have proposed a new, more powerful dynamic model for tracking multiple people in complex scenarios. The LTA model is not dependent on any specific tracker or scene, it merely needs the subjects to reside in a space that allows one to calculate metric distances.

The LTA model takes into account both simple scene information in the form of destinations or desired directions, and interactions between different targets. As it operates in world coordinates, the model can be trained offline on training sequences, and then applied elsewhere. We have shown experimentally that the model yields better predictions, and consistently improves tracking performance compared to dynamic models which discregard social interaction. The improvement comes at negligible computational cost (less than 10 ms for a frame with 15 subjects).

We draw attention to an additional lesson learned from the study: a person's destination is valuable information and should always be used. While this finding is by no means new, *e.g.* [14, 12], we emphasize that it is true even when the destinations are incomplete or inaccurate. We have shown that even roughly guessed target directions help to make more meaningful predictions. This is particularly interesting for the case of mobile cameras, where the destination cannot be learned from continuous observation.

In the present state, we do not model groups of people walking together. This would be possible by an extension to the energy potential. A further interesting direction is the stochastic application of the proposed energy functional.

Acknowledgments—This project has been funded in parts by Toyota Motor Corporation, and by EU projects DIRAC (IST-027787) and EUROPA (ICT-2008-231888). We are grateful to Paolo Urbano for inspiring discussions and to Vittorio Ferrari for suggesting the paper title.

References

[1] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*, 2008.

[2] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR'08*.

[3] G. Antonini, S. V. Martinez, M. Bierlaire, and J. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *IJCV*, 69:159–180, 2006.

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008.

[5] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009.

[6] G. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *CVPR*, 2006.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] A. Ess, B. Leibe, K. Schindler, and L. van Gool. A mobile vision system for robust multi-person tracking. In *CVPR'08*.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[10] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR*, 2006.

[11] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995.

[12] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV'08*.

[13] A. Johansson, D. Helbing, and P. K. Shukla. Specification of a microscopic pedestrian model by evolutionary adjustment to video tracking data. *Advances in Complex Systems*, 10(2):271–288, 2007.

[14] R. Kaucic, A. G. Perera, G. Brooksby, J. Kaufhold, and A. Hoogs. A unified framework for tracking through occlusions and across sensor gaps. In *CVPR*, 2005.

[15] F. Klügl and G. Rindsfüser. Large-scale agent-based pedestrian simulation. In *MATES '07: Proc. of the 5th German Conference on Multiagent Systems Technology*, 2007.

[16] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool. Coupled detection and tracking from static cameras and moving vehicles. *IEEE TPAMI*, 30(10):1683–1698, 2008.

[17] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *EUROGRAPHICS*, 2007.

[18] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009.

[19] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, 2004.

[20] A. Penn and A. Turner. Space syntax based agent simulation. In *PED*, 2002.

[21] A. Schadschneider. Cellular automaton approach to pedestrian dynamics—theory. In *PED*. 2001.

[22] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet part detectors. *IJCV*, 75(2):247–266, 2007.

[23] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008.

[24] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *CVPR*, 2004.