Wrong Turn – No Dead End: a Stochastic Pedestrian Motion Model

Stefano Pellegrini[1] Andreas Ess[1] Marko Tanaskovic[1] Luc Van Gool[1,2]

[1]Computer Vision Laboratory
ETH Zurich
{stefpell,aess,vangool}@vision.ee.ethz.ch

[2]ESAT-PSI / IBBT
KU Leuven
tmarko@student.ethz.ch

Abstract

This paper addresses the use of social behavior models for the prediction of a pedestrian's future motion. Recently, such models have been shown to outperform simple constant velocity models in cases where data association becomes ambiguous, e.g. *in case of occlusion, bad image quality, or low frame rates. However, to account for the multiple alternatives a pedestrian can choose from, one has to go beyond the currently available deterministic models. To this end, we propose a stochastic extension of a recently proposed simulation-based motion model. This new instantiation can cater for the possible behaviors in an entire scene in a multi-hypothesis approach, using a principled modeling of uncertainties. In a set of experiments for prediction and template-based tracking, we compare it to a deterministic instantiation and investigate the general value of using an advanced motion prior in tracking.*

1. Introduction

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 been used for quite some time in other communities, such as in Computer Graphics [6, 7] or Social Science [5, 10, 12]. Researchers in Computer Vision, on the other hand, have often resorted to using a simple constant velocity assumption for pedestrians, especially in applications such as tracking. Recently however, more advanced models, often inspired by social simulations, received considerable attention. Antonini *et al*. [2] were one of the first to use a behavioral motion prior in a tracker, using a Discrete Choice Model to select the next position for each pedestrian. Ali and Shah [1] use scene-specific "floor field" models to make tracking in extremely crowded situations tractable. In our previous work [9] we built the motion prior on the assumption that each subject predicts the other pedestrians' trajectories with a simple linear extrapolation, calculating the next velocity based on this prediction. [13] focus on learning the parameters of a space-continuous time-discrete model that is optimized with a gradient-descent technique. Outside the area of tracking, Mehran *et al*. [8] use the social force model to detect abnormal behavior in crowded scenes.

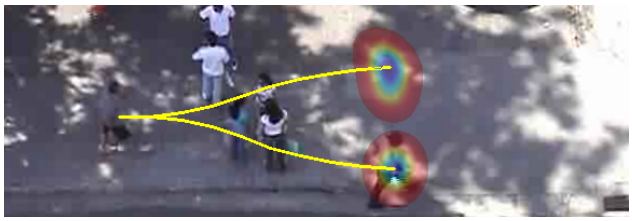


Figure 1. When moving through a scene, a person takes a variety of factors into account, such as steering clear of other people. In many cases, the prediction of the motion cannot be well described by a deterministic algorithm: in the above example the pedestrian on the left hand side could either evade the group by going on its left or right side, as indicated by the yellow lines. We therefore propose a stochastic, simulation-based motion model that can deal with the uncertain future motion of a pedestrian.

When scene-specific knowledge is not available, a microscopic model (handling pedestrians separately) is usually preferred over a macroscopic one (focusing on a crowd's behavior, rather than its individual members). Microscopic models usually account for interactions among individuals, destinations, and desired velocities. However, accurately modeling a pedestrian's future path in a deterministic way is almost impossible: on the one hand, the observed information is incomplete, either because it is invisible to the camera (but visible to the pedestrian in the scene), or because it is part of a pedestrian's individual preferences (some people like to walk in the shade, others do not). On the other hand, model complexity is limited by computational power. Instead of trying to model more and more factors, we therefore propose to use a deterministic motion model, in this case Linear Trajectory Avoidance (LTA) [9], and extend it to allow multiple hypotheses. The ensuing novel stochastic formulation can then handle such unaccounted factors in a probabilistic way.

978-1-4244-7030-3/10/$26.00 ©2010 IEEE

In this paper, we therefore introduce a stochastic extension of LTA, in order to make the motion prior more robust to this kind of failures, Fig. 1. We term the ensuing new model stochastic LTA, or sLTA. sLTA uses the same energy potential formulation as the original LTA model, but in a Gibbs form to turn the potential into a probability. Multiple modes (*i.e.*, alternatives in choosing a path) in the ensuing distribution are then approximated by a mixture of Gaussians, each giving rise to another set of possible future world state. For each further time step, this recursively yields a tree of possible future locations, with an uncertainty defined through mathematically sound error propagation.

One specific question that is then addressed is the usability of the motion model for tracking. In [9], it has already been shown that a motion prior has better predictive power than linear extrapolation and that a tracker can benefit from its use in situations where the observation is unreliable (*e.g.*, during occlusions). In this paper, we investigate this issue further by conducting a set of systematic experiments using an appearance-based tracker.

The paper is structured as follows. After briefly introducing the basic motion model (LTA) in Sec. 2, we introduce its novel stochastic extension in Sec. 3. Experiments are presented in Sec. 4, before the paper is concluded in Sec. 5.

2. LTA, a Pedestrian Motion Model

LTA (linear trajectory avoidance) is a simulation-based motion model [9] . It predicts a pedestrian's current velocity[1] $\mathbf{v}_i^t$ based on the previous positions $\mathbf{x}_j^{t-1}$ and the velocities $\mathbf{v}_j^{t-1}$ of all the N pedestrians j in the scene, as well as on the pedestrians' desired destinations $\mathbf{r}_j$ and desired speeds u_j, and on the static obstacles represented by an obstacle map $\mathbf{I}$. Similar to other simulation-based motion models, LTA assumes that pedestrians interact with each other, trying to keep a certain distance while walking to a desired destination with a certain speed. These *social* factors are included into an energy potential

$$E(\mathbf{v}_i^t|\mathbf{S}^{t-1}, \mathbf{U}, \mathbf{R}, \mathbf{I}) \ , \tag{1}$$

where

$$\mathbf{S}^{t-1} = [\mathbf{x}_0^{t-1}, \mathbf{v}_0^{t-1}, \ \dots \ \mathbf{x}_N^{t-1}, \mathbf{v}_N^{t-1}] \tag{2}$$

$$\mathbf{U} = [u_0, \ \dots \ u_N] \tag{3}$$

$$\mathbf{R} = [\mathbf{r}_0, \ \dots \ \mathbf{r}_N] \ . \tag{4}$$

Taking the minimum of this potential yields a pedestrian's next *desired* velocity that is used to update the model.[2] The procedure is repeated for each pedestrian independently in parallel, and illustrated in Fig. 2. While in [9] we show that taking into account these social factors improves prediction and object tracking performance compared to a constant velocity model, there are two main problems with the formulation: firstly, deterministically choosing the minimum of the energy function cannot account for the multiple options a pedestrian can choose from when walking onto other people. Therefore, a deterministic instantiation of the model sometimes commits big errors when avoiding oncoming pedestrians on the wrong side. Secondly, uncertainty propagation is handled only empirically.

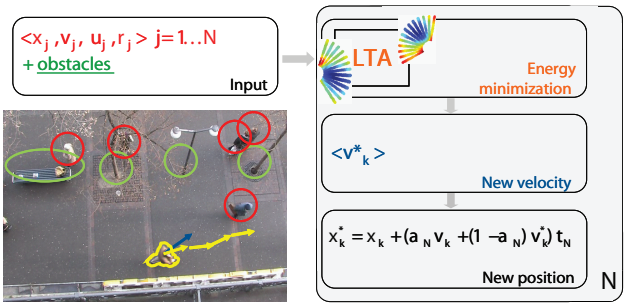


Figure 2. Principle of LTA: given the state of the current image (red: pedestrians, green: obstacles), every pedestrian is simulated in turn, assuming a simple path-planning behavior of each individual. After finding the most probable velocity for each pedestrian (in a deterministic fashion), their position is updated in parallel. In this paper, we extend the deterministic behavior with a stochastic one, to account for the uncertainty of a person's future motion.

To this end, we propose an improved version of LTA, extending it to handle the uncertainty in a still approximate but more principled manner. Above all, rather than representing the output of the algorithm with only a single value, we will extend it to handling multiple choices for each pedestrian.

3. Stochastic LTA

To account for the uncertain future motion of a pedestrian, we extend LTA in a multi-hypothesis fashion. Based on the energy potential from Eq. (1), we define the posterior probability of a pedestrian's velocity $p(\mathbf{v}_i^t|\mathbf{S}^{t-1}, \mathbf{U}, \mathbf{R}, \mathbf{I})$ as a Gibbs potential,[3] for each pedestrian as

$$p(\mathbf{v}_i^t|\mathbf{S}^{t-1}) = Z^{-1} e^{-\omega E(\mathbf{v}_i^t|\mathbf{S}^{t-1})} \ , \tag{5}$$

[1] As in physics, we use the term velocity for a two-dimensional motion vector, as opposed to the scalar value speed.

[2] In the original model, the desired velocity is linearly filtered for smoothness (see Eq. 14 in [9]). In this paper, we use an equivalent energy potential that includes already the same smoothing, by introducing a simple coordinate transformation:

$$E(v^t|\mathbf{S}^{t-1}, \mathbf{U}, \mathbf{R}, \mathbf{I}) = E_{LTA}(\frac{v^t - \alpha * v^{t-1}}{1-\alpha})$$

where E_{LTA} is the formulation of the energy given in [9]. Note that this is an entirely equivalent formulation, but has the advantage of being more compact.

[3] To reduce notational complexity, we will omit the dependency on $\mathbf{U}, \mathbf{R}, \mathbf{I}$ in the rest of the paper.

where Z is a normalization constant and ω is a free parameter that will be discussed later. We now assume that a pedestrian, at each time step t, makes a decision for his next velocity v_t based on its past observations of the environment. As opposed to standard LTA, we allow multiple alternatives, or hypotheses. Therefore, rather than working with Eq. (5), we fit a mixture of Gaussians

$$p(\mathbf{v}_i^t|\mathbf{S}^{t-1}) \approx \sum_k^K w_k \mathcal{N}(\mathbf{v}_i^t|\mathbf{v}^t_{LTA(k|\mathbf{S}^{t-1})}; \Psi^t_{LTA(k|\mathbf{S}^{t-1})}) \ , \tag{6}$$

where the LTA subscript indicates that the quantity is estimated from LTA. This mixture could be fit with standard methods such as Expectation Maximization or iterative function fitting techniques. However, to keep the system applicable to real-time scenarios, we opt to use the following heuristic to estimate the mixture parameters:

1. Discretize the distribution of Eq. (5). The number of components K of the mixture is decided by counting the local maxima in the discretized distribution.
2. Run a (BFGS) minimization for each mode to refine the mode estimate. These mode estimates are assumed to be the locations of the means $\mathbf{v}^t_{LTA(k|\mathbf{S}^{t-1})}$ of the mixture components.
3. Label the (negative) basins of attraction of each maxima in the discretized grid and use this clustering to estimate the covariances $\Psi^t_{LTA(k)}$.
4. The weight w_k of each mode is computed for each component independently, by setting the k^{th} component's mode of the mixture equal to the energy at that point

$$w_k = \frac{exp(-\omega E(\mathbf{v}^t_{LTA(k|\mathbf{S}^{t-1})}))/Z}{\mathcal{N}(\mathbf{v}^t_{LTA(k|\mathbf{S}^{t-1})}; \mathbf{v}^t_{LTA(k|\mathbf{S}^{t-1})}, \Psi^t_{LTA(k)})}. \tag{7}$$

These weights are finally normalized so that their sum is one (therefore, the equality in Eq. (7) does not necessarily hold anymore (see also Fig. 3).

This is obviously a rough estimate of the parameters, that will become worse the less the Gaussians are separated. Nevertheless, it turned out to be sufficient for our purposes (see Fig. 3 for an example fit). In Sec. 5, we explain why the algorithm is robust in this respect.

Let us now assume that subject i's position at time $t-1$ is distributed as $\mathcal{N}(\mathbf{x}_i^{t-1}|\boldsymbol{\mu}_i^{t-1}; \boldsymbol{\Sigma}_i^{t-1})$, that is a single Gaussian. Assuming a linear process of the form

$$\mathbf{x}_i^{t-1} = \mathbf{x}_i^{t-1} + \Delta_t \mathbf{v}_i^{t-1} + \gamma \text{ with } \gamma \sim (\mathbf{0}; \Gamma) \tag{8}$$

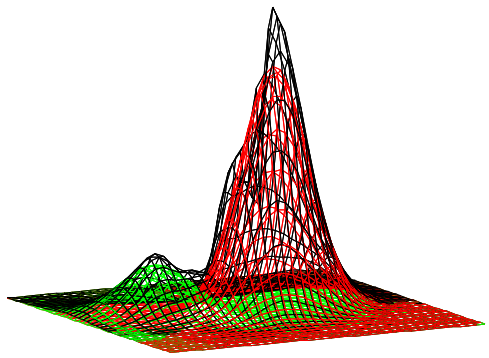


Figure 3. The energy potential is brought into an analytical form by fitting a mixture of Gaussians using a fast approximative method (see text).

and using Eq. (6), we can write

$$p(\mathbf{x}_i^t) = \int p(\mathbf{x}_i^t, \mathbf{S}^{t-1}) d\mathbf{S}^{t-1} \tag{9}$$

$$= \sum_n^K w_k \mathcal{N}(\mathbf{x}_i^t|\boldsymbol{\mu}^t_{i_k}; \boldsymbol{\Sigma}^t_{i_k}) \ , \tag{10}$$

where for subject i

$$\boldsymbol{\mu}_k^t = \boldsymbol{\mu}_k^{t-1} + \Delta_t \mathbf{v}^t_{LTA(k)} \tag{11}$$

$$\boldsymbol{\Sigma}_k^t = \Gamma + \boldsymbol{\Sigma}_k^{t-1} + \Delta_t^2 \Psi^t_{LTA(k)} \ , \tag{12}$$

(see App. A for more details about the derivation). In order to make use of Eq. (10) for one subject we need, however, to specify the positions and velocities of all other pedestrians, i.e. $\mathbf{S}^{t-1}$. At each time step though, we are only provided with a distribution of the pedestrian position that, as shown in Eq. (10), has the form of a Gaussian mixture.

Using only the mean of the distribution is not a suitable statistic for a mixture of Gaussians. Alternatively, selecting only the best mode would result in a deterministic model with only a single hypothesis. Instead, we start at time $t = 0$ with a single gaussian per pedestrian and we compute the probability for each of the $M = \prod_i^N K_i$ possible combinations m of mixtures modes at time t as $\prod_i^N w_{i_m}$, where i_m is the index of the alternative that subject i chooses in combination m. We assume that these combinations are independent from each other (see App. A). In each combination m, a subject is represented by its $i_m - th$ mode, that is a single Gaussian. By using Eq. (10) now, we can use the values of the modes that participate in the combination as the position of the pedestrians, while we can estimate the velocity by backtracking the last position in the pedestrian's past trajectory. Repeating this process at each time step, for each combination, obviously leads to a combinatorial explosion. Each combination indeed splits into new

ones. To prevent this from happening, we limit the maximum number of combinations to a value $\hat{M}$. If the splitting process at a certain time step generates more than $\hat{M}$ combinations, the most likely $\hat{M}$ are used, while the others are discarded. Further, we only allow the splitting into multiple modes of Eq. (10) when the probability of the combination is $> \epsilon = 0.1$. Since the probability of the combinations decreases with time because of the splitting, at a certain point the splitting ceases.

Note that in the special case when $\hat{M} = 1$, the model is deterministic and almost the same as the original LTA . The main difference is that in the original LTA, the next velocity v^* was computed with a gradient descent over the energy potential E_{LTA}, while now the heuristic just described is used.

Note also that this general approach of handling multiple possible world states is conceptually similar to multi-hypothesis tracking [11], in which each world corresponds to a possible data association between trajectories and observations.

3.1. Why not a Particle Filter Framework?

Eq. (5) could be easily used in a particle filter framework as a propagation function (see Fig. 4). It is reasonable to expect that the results, for a sufficient number of particles, are more accurate than those obtained with an approximation by a mixture of Gaussians. However, there are at least two reasons why to refrain from taking this approach.

The first reason is related to computational requirements. Since we want to represent the interactions between subjects, the state space cannot be easily factored into independent particle filters. The state should rather be represented jointly by the positions and velocities of all the subjects. With the ensuing rapidly growing state dimension, the number of particles increases exponentially. For each particle, the basic LTA procedure should be evaluated for each subject, which is computationally prohibitive. In contrast, in our formulation the LTA procedure is only invoked for each mode of the mixture.

Even if a particle filter would be computationally feasible, we believe that the commonly used resampling stage [3] introduces a higher logic that we assume a pedestrian to not have in the LTA model: if a mode of the sampled distribution happens to die out at some point, *e.g.* due to higher likelihood of the other modes in the resampling stage, the history of the particles belonging to the cloud until that point is meaningless. Once an alternative has been created, it cannot cease to exist simply because, a posteriori, other alternatives are more suited. This would imply that pedestrians predict their complete possible future trajectories in advance, even with information that is unavailable to them at present, and then choose the feasible ones. This assumption is not part of the LTA model and also does not seem to be realistic.

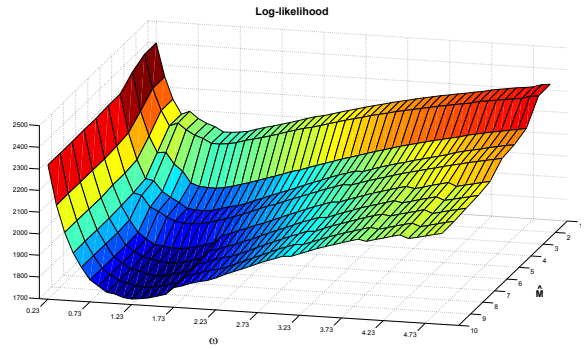


Figure 5. Log-likelihood of number of combinations $\hat{M}$ and free parameter ω. Increasing the number of combinations always improves the result.

3.2. Training

For training the underlying LTA, we employ the procedure described in [9], using the training set provided by the authors. The parameters for the stochastic version then remain the same, we inspect the effect of the remaining free parameters $(\hat{M}, \omega)$ in the experiments section.

4. Experiments

In the following, we first evaluate the sLTA model by itself, comparing its prediction capabilities for different parameter settings. We then show its application in a tracking experiment, highlighting the importance of a good motion model in data association. For these experiments, we use annotated data provided by the authors of [6]. The video shows part of a shopping street from an oblique view. A homography from image to ground plane was estimated from four manually clicked points on the footpath to transfer image to world coordinates. Standing and erratically moving people were marked; for these, a simple extrapolation is used. As destinations we chose two points far outside the left and right image borders, which holds for most subjects. Static obstacles (*i.e.*, the building and the parked car) were also annotated.

4.1. Prediction

To test the prediction capabilities of our model, we evaluate on a subsequence of about 3 minutes @ 2.5 *fps*, containing 86 trajectories annotated with splines. We simulate all the subjects in parallel. Note that this is different from the prediction experiment in [9], where each subject was simulated in turn, while using the ground truth positions and velocities of all the others. Starting one simulation every 1.2 seconds with a prediction horizon of 4.8 seconds yields $\approx$ 200 simulations. To highlight the importance of using multiple modes, as well as the effect of the parameter ω, we run multiple simulations over all sub-

Figure 4. Particle filter experiment: when simulating a person (yellow circle) given the other people (green circles) using a particle filter embodiment of the model, multiple modes (red particles) form naturally. While both options of steering clear of the oncoming persons are found, such a solution is computationally prohibitive (see text).

jects, varying both the maximum number of combinations $\hat{M}$, as well as the free parameter ω. For each simulation, we report the log-likelihood $\log p(GT|\hat{M}, \omega)$ of $\hat{M}$ and ω based on the ground truth trajectories GT, Fig. 5. As can be seen, increasing the number of combinations, and therefore of modes, always improves the prediction result, irrespective of the chosen ω: this indicates that even with multiple modes, the model is conservative enough as not too allow completely improbable predictions. The parameter ω relates to how certain each hypothesis is. When ω is zero, the probability is uniform, while for bigger and bigger values of ω, the uncertainty around each mode decreases. Fig. 5 shows a small yet interesting positive correlation between the value of ω and $\hat{M}$. This can be interpreted saying that when increasing the number of combinations, less uncertainty per mode is *allowed*.

Some example images when using 10 modes are shown in Fig. 6. Red lines indicate the ground truth, yellow lines indicate the predicted path of a person, blue circles correspond to the standard deviation of the fitted Gaussians at the respective end positions. Green lines indicate the linear extrapolations of people that are standing or moving erratically, white boxes the set of used obstacle points. Please note that the model operates in ground-plane coordinates, hence all drawings correspond to people's feet in the image. For each image, we show the final image after 4 s of extrapolation. As can be seen, the model manages to find the correct extrapolation for almost all persons in one of its modes, while keeping the number of modes at a minimum. Multiple possibilities can be especially seen when people are walking towards other groups of people, *e.g.*, in the top right and lower left image.

In the deterministic setting ($\hat{M} = 1$), extrapolations in easy situations remain the same (Fig. 7, left; these images correspond to the top line of Fig. 6). In more difficult situations, only the stronger mode remains, which can either be correct (middle) or wrong (right). Thus, from a prediction point of view, it is indeed beneficial to use multiple modes in a stochastic fashion, as suggested by this paper.

Finally, Fig. 8 shows some typical failures of the model. These are not all failures in the hard sense, as the stochastic options often also includes the correct solution: in the first image, the model splits too much because it is unsure what to do with two persons walking with each other in a group, but slightly changing positions to each other. It splits, but keeps the correct hypothesis. In the middle image, another person is wrongly extrapolated (green line in middle of image), causing a split, but the correct hypothesis is also kept. In the last image, the lower extrapolation is wrong, with the correct solution (going above the standing group) not identified: this is a special case of the first case, where two people walking in a group feel repulsion rather than staying together. Implementing the notion of groups would alleviate such problems.

4.2. Tracking

To explore the effect of a stochastic motion prior on tracking performance, we present the following experiment: for each person, and for increasing time horizons, we perform an NCC-based template matching between a subject in a reference frame and its possible location in a later frame. The chosen motion model defines the search radius for the matching; the solution is found as the peak NCC-response, weighted by the motion models' uncertainty. The error in distance between this solution and the ground truth is accumulated for all persons and by starting the tracking every 1.2 seconds. As the model is trained in steps of 0.4 seconds (10 frames), we also keep this spacing for the experiment.

This experiment should highlight the advantage of a good motion model: a correct search region should prevent the tracker from drifting by guiding the data association. Instead of including the model into a complicated tracker, where many side-effects can influence the result, we therefore keep the experiment as simple as possible to see the real merit of a motion model.

We specifically compare a simple Brownian motion model with a constant velocity one, as well as different instantiations of sLTA. For the experiment, we use templates of 30×30 pixels on people's head positions. As an additional baseline, we use an adaptive tracker based on online boosting [4] that uses *all* intermediate frames (as opposed to steps of 10 frames). In the given sequences, purely appearance-based matching is especially tricky due to low contrast, cast shadows, and interlacing and compression artifacts. The motion model uncertainty is chosen as follows: for the Brownian model, the uncertainty is assumed uniform

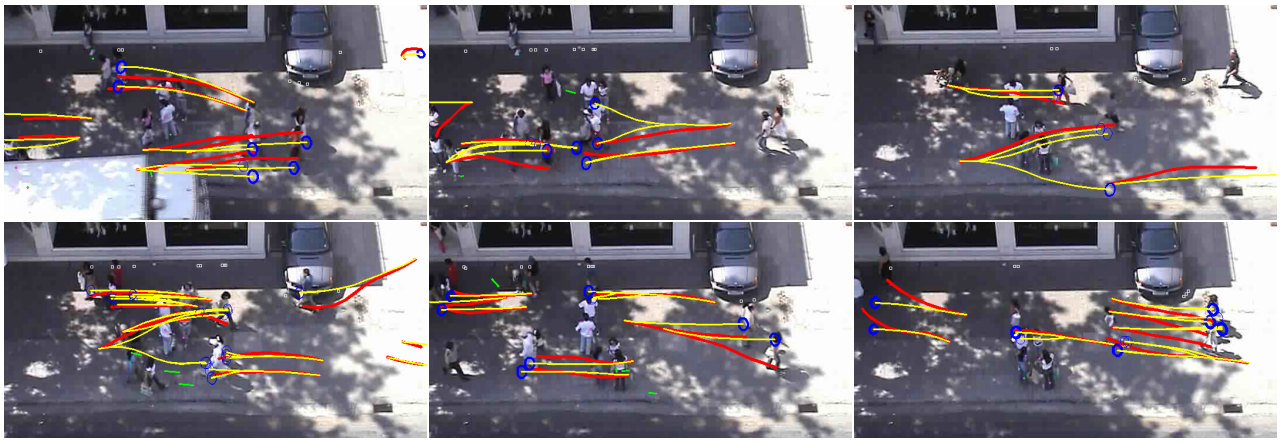


Figure 6. Example extrapolations. The possible paths for a given person are shown in yellow, with blue circles indicating the σ-confidence of the fitted Gaussians. Note that the model operates in ground-plane coordinates, the lines and circles thus correspond to people's feet. Also note that all the subjects are simulated in parallel.

Figure 7. Extrapolations when just using one mode, corresponding to a deterministic model. (see text)

in the search region (which is bounded by a statistic on the maximum walking speed); for the constant-velocity one, we use a single Gaussian centered around the prediction (we plot results for two choices of the uncertainty); for sLTA, the mixture of Gaussians as introduced above is used. In this systematic experiment, velocities are inferred from the past frame's ground-truth. While this does not reflect the actual tracking application, it still allows for a fair comparison between the different models, and their (ideal) influence on appearance-based tracking.

Fig. 9 (a) plots the mean error in meters for all approaches. We furthermore report the number of actual tracking errors (deviation from ground truth > 0.5 m) in Fig. 9 (b). For increasing frame gaps, an uninformed motion model makes tracking virtually impossible ("Brownian", mean error not plotted in (a) due to large error). For small time horizons, the result of a constant velocity model ("Const. vel.") is virtually the same as with any more advanced model, as small motions can be sufficiently approximated by a linear extrapolation. However, for increasing time horizons, the positive effect of sLTA becomes more pronounced. This is also in line with other researchers' results [9], who mainly observed an effect of a strong motion model in cases of missing data, *e.g.* due to occlusion.

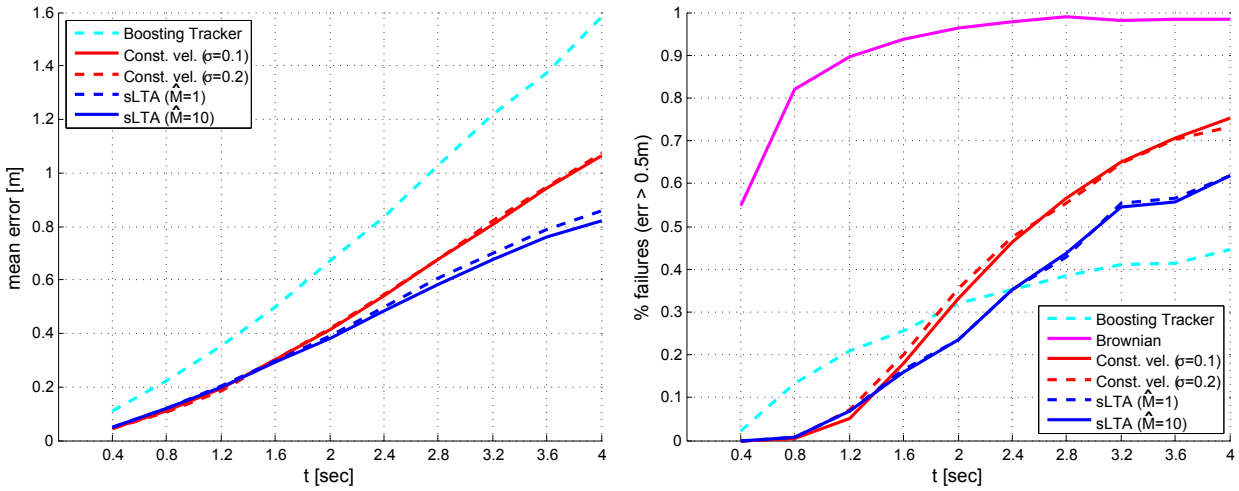
As an additional baseline, we show the result of purely appearance-based tracker, which uses *all* available intermediate frames while learning the model of the appearance ("Boosting Tracker"). Using all available data from the image produces fewer hard failures, still, the high mean error indicates that when the tracker starts drifting, it's totally lost. We therefore believe a strong motion model to be important for tracking.

Accounting for a pedestrian's future motion in a probabilistic manner, *i.e.*, using $\hat{M} = 10$ instead of $\hat{M} = 1$, does not seem to have a considerable effect on tracking performance: both the mean error and the fraction of tracking errors seems to only improve slightly when allowing multiple modes. The important thing to note here is that in the presented sequence, there is only a limited number of "splittings" in general, and only in a fraction of these, the deterministic model chooses the wrong mode. While the effect thus seems limited, this still means that in such cases, the tracker would fail and lose an object for multiple seconds, searching in the wrong location. Employing a stochastic model therefore definitely helps in extreme situations, which can also be expected more frequently in more crowded scenarios.

5. Discussion and Conclusions

This paper presented a sound stochastic extension of an existing, simulation-based motion model for pedestrians. The novel probabilistic formulation is based on using the

Figure 8. Typical failures of the sLTA: (Left, middle) unnecessary splittings can occur due to other wrong extrapolations, but are handled in the multi-hypothesis framework. (Right) without the knowledge of people walking in groups, wrong extrapolations can occur. (see text for details)



(a) (b)

Figure 9. (a) Mean error (in meters) of tracking using different motion models, for increasing frame gaps. (b) Number of tracking failures (error > 0.5 m) using different motion models, for increasing frame gaps.

original energy function as a Gibbs potential. Then, by using a multi-hypothesis approach with mathematically sound uncertainty propagation, a set of possible future world states is obtained. To achieve a good compromise between accuracy and tractability, we fit a gaussian mixture model to the Gibbs potential. Although the fitting is rather approximate, we found it to work well in our experiments. This is due to the fact that the actual choices of pedestrians seem to be limited to one or two, for each timestep. Therefore the potential will have only one or two modes. Furthermore, the modes corresponding to alternatives of a choice for a pedestrian, tend to separate apart with time. This allows us to *wait* for the modes to be well separated before fitting the mixture (when their distance is below an empirical threshold, we group them and consider them as a single mode).

In our prediction experiments, we showed that the log-likelihood of the prediction increases considerably as we go from a deterministic instantiation to a stochastic one.

For tracking, a clear advantage over simpler motion models was demonstrated, the effect of a stochastic model is however not as pronounced as expected. While more complicated scenes would probably show an advantage of using a probabilistic formulation, this difference is only present at higher frame gaps, which could be *e.g.* due to occlusion. Generally, it thus seems that the prediction would be more suited to tasks in, *e.g.*, robot navigation, where safety is a crucial issue.

Future work will therefore study the application of the model to robot navigation. Furthermore, we plan to explore the grouping behavior between pedestrians. As we showed in a few examples, the knowledge of whether a person belongs to a group can influence her/his motion planning.

Acknowledgments—This project has been funded in parts by Toyota Motor Corporation, and by EU projects DIRAC (IST-027787) and EUROPA (ICT-2008-231888).

A. Derivation of Empiric Covariance

The derivation of $p(\mathbf{x}_i^t, \mathbf{v}_i^t)$, and $p(\mathbf{x}_i^t)$ for a given pedestrian i is shown in Tab. 1 We describe hereafter the simplifying assumptions we make in order to obtain a tractable and real-time capable solution. First, in Eq. (14), we assume that the state space S^{t-1} can be partitioned into M subsets S_m^{t-1}. Each of the subsets encode a possible combination

$$
\begin{aligned}
p(\mathbf{x}_i^t, \mathbf{v}_i^t) &= \int p(\mathbf{x}_i^t, \mathbf{v}_i^t | S^{t-1}) p(S^{t-1}) dS^{t-1} && (13)\\
&= \sum_m^M \int p(x_i^t, \mathbf{v}_i^t | S_m^{t-1}) p(S_m^{t-1}) dS_m^{t-1} && (14)\\
&= \sum_m^M \int \mathcal{N}(\mathbf{x}_i^t | \mathbf{x}_{i_m}^{t-1} + \Delta_t \mathbf{v}_i^t; \Gamma) \sum_{k_m}^{K_m} w_{k_m} \mathcal{N}(\mathbf{v}_i^t | \mathbf{v}_{LTA(k_m)}^t; \Psi_{LTA(k_m)}^t) \prod_{j_m} p(\mathbf{x}_{j_m}^{t-1}, \mathbf{v}_{j_m}^{t-1}) d\mathbf{v}_{j_m}^{t-1} d\mathbf{x}_{j_m}^{t-1} && (15)\\
&= \sum_m^M w_m \sum_{k_m}^{K_m} w_{k_m} \int \mathcal{N}(\mathbf{x}_i^t | \mathbf{x}_{i_m}^{t-1} + \Delta_t \mathbf{v}_i^t; \Gamma) \mathcal{N}(\mathbf{v}_i^t | \mathbf{v}_{LTA(k_m)}^t; \Psi_{LTA(k_m)}^t) \mathcal{N}(\mathbf{x}_{i_m}^{t-1} | \mu_{i_m}^{t-1}; \mathbf{\Sigma}_{i_m}^{t-1}) dx_{j_m}^{t-1} && (16)\\
&= \sum_m^M w_m \sum_{k_m}^{K_m} w_{k_m} \mathcal{N}(\mathbf{x}_i^t | \mu_{i_m}^{t-1} + \Delta_t \mathbf{v}_i^t; \Gamma + \mathbf{\Sigma}_{i_m}^{t-1}) \mathcal{N}(\mathbf{v}_i^t | \mathbf{v}_{LTA(k_m)}^t; \Psi_{LTA(k_m)}^t) && (17)\\
&= \sum_c^C w_c \mathcal{N}(\mathbf{x}_i^t | \mu_{i_c}^{t-1} + \Delta_t \mathbf{v}_i^t; \Gamma + \mathbf{\Sigma}_{i_c}^{t-1}) \mathcal{N}(\mathbf{v}_i^t | \mathbf{v}_{LTA(c)}^t; \Psi_{LTA(c)}^t) && (18)\\
p(\mathbf{x}_i^t) &= \int p(\mathbf{x}_i^t, \mathbf{v}_i^t) d\mathbf{v}_i^t && (19)\\
&= \sum_c^C w_c \mathcal{N}(\mathbf{x}_i^t | \underbrace{\boldsymbol{\mu}_{i_c}^{t-1} + \Delta_t \mathbf{v}_{LTA(c)}^t}_{\boldsymbol{\mu}_{i_c}^t}; \underbrace{\Gamma + \mathbf{\Sigma}_{i_c}^{t-1} + \Delta_t^2 \Psi_{LTA(c)}^t}_{\mathbf{\Sigma}_{i_c}^t}) && (20)
\end{aligned}
$$

Table 1. Derivation of $p(\mathbf{x}_i^t, \mathbf{v}_i^t)$ and $p(\mathbf{x}_i^t)$

for the state of the pedestrians at time $t-1$, each state being described with a single Gaussian distribution (a multivariate one, for position and velocity). Furthermore, we assume that these combinations are well separated from each other, so that we can consider them independently and not affecting each other (imagine the typical case when being able to avoid a pedestrian on either left or right side). In Eq. (15), we therefore assume that the state space S^{t-1} can be factorized into individual factors, each depending only on the position and velocity of a single pedestrian. Eq. (15) also introduces the LTA model and consequently the approximation by a mixture of Gaussians (see Eq. (6)). Here we use k_m to indicate the $k-th$ alternative of subject i starting from combination m, therefore dropping index i to reduce notational clutter. In Eq. (16), all the variables but $x_{j_m}^{t-1}$ are integrated out. Here, w_m results from the multiplication of the weights the $p(\mathbf{x}_{j_m}^{t-1}, \mathbf{v}_{j_m}^{t-1})$ components were carrying. In both Eq. (17) and Eq. (20) the Gaussian marginalization is applied. In Eq. (18), for notational simplicity, a new index c is introduced (we omit the mapping $\mathrm{Map}(k_m, m) \rightarrow c$.

References

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

[2] 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. 1

[3] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans on Sig Proc*, 2002. 4

[4] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR*, 2006. 5

[5] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical Review*, 51(5), 1995. 1

[6] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *EUROGRAPHICS*, 2007. 1, 4

[7] Massive Software. Massive, 2010. 1

[8] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using Social Force model. In *CVPR*, 2009. 1

[9] S. Pellegrini, A. Ess, K. Schindler, and L. V. Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 1, 2, 4, 6

[10] A. Penn and A. Turner. Space syntax based agent simulation. In *Pedestrian and Evacuation dynamics*, 2002. 1

[11] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Trans. on Automatic Control*, 24(6):843–854, 1979. 4

[12] A. Schadschneider. Cellular automaton approach to pedestrian dynamics - theory. In *PED*, 2001. 1

[13] P. Scovanner and M. Tappen. Learning pedestrian dynamics from the real world. In *ICCV*, 2009. 1