BEHAVIORAL FILTERING OF HUMAN TRAJECTORIES FOR AUTOMATIC-MULTI-TRACK INITIATION

Gianluca Antonini, Santiago Venegas, Michel Bierlaire and Jean-Philippe Thiran

Swiss Federal Institute of Technology (EPFL)
Signal Processing Institute (ITS) and Operation Research Chair (ROSO)
CH-1015 Lausanne, Switzerland
home page: http://ltswww.epfl.ch
{Gianluca.Antonini,Santiago.Venegas,Michel.Bierlaire,JP.Thiran}@epfl.ch

ID 79

ABSTRACT

In this paper we describe a new approach for the multi-track initiation problem. We propose an extensive use of the top-view reconstruction of the scene to solve the detection step in tracking pedestrians. We leave a large set of starting hypothetical moving objects free to evolve in the scene for a certain number of frames. The number of trajectories is pre-filtered using distance and direction constraints on a one-step movement of pedestrians. The output trajectories from pre-filtering step are then filtered using a discrete choice model for pedestrian behavior, calibrated on real data. The results show that is possible to use this technique to perform multitarget tracking in real situations. We particularly focus on an application in the context of automatic video surveillance.

1. INTRODUCTION

Automated visual surveillance has a variety of potential applications. Detection and tracking of suspicious moving objects (pedestrians or vehicles) represent a fundamental task in this context. There are two main approaches in the field of object tracking algorithms. A class of algorithms uses differential techniques based on the assumption that the object does not move much from one frame to the next and employing a local search around the previous object position to locate the moving object on the next frame ([1]). The drawback of these techniques is that large displacements are not supported.
Another class of algorithms makes use of feature detection ([2]). Multiple hypothesis tracker (particle filters) ranks the possible hypothesis in order of likelihood. It supposes that the correct hypothesis is retained. In this case, tracking filters have been proposed using the state-space based modeling that defines each static posture or position as a state and describes a motion sequence by the composition of these states with some transitional probabilities ([3]).
In both approaches the data processing includes track initiation, maintenance, and termination. However, the proposed techniques perform efficiently to trace the mouvement of few moving objects. The operational efficiency decreases dramatically when the number of targets increases. Moreover, they rarely address the problem of automatic target detection, i.e. of track initiation. Finally, all these techniques strongly depend on the image quality.
The detection problem or track initiation is the fundamental operation of a tracking system. This operation distinguishes objects of interest which have to be tracked ([1]). Traditionally the hypothetical foreground regions are detected and automatic target recognition is performed to reject them or to start a track. In order to simplify the target recognition task, the video camera must be ideally placed at the top of the scene to avoid projection and occlusion between objects. Search results, for nowdays, have not yet demonstrated a segmentation method able to identify the objects that overlap each other within video frames.
To face these problems, we think a good discriminant to detect target regions is their dynamic behavior. In this paper we use the top-view plan to formulate behavioral constraints that we use to detect individuals on the image plan, analysing their trajectories over a certain time period T. The top-view plan is a reconstruction of the position of each region in the real scene, obtained by a calibrated camera. In the case of pedestrian trajectory this reconstruction gives the position of each pedestrian on the top-view plan of the scene and not its position projected on the image plan. The method is robust against bad image quality and illumination conditions in cluttered environments.
The paper is organized as follows. In section 2 we give a

general overview of our system. In section 3 we address the problem of the generation of the starting hypothetical moving object positions. In section 4 we discute and analyse the behavioral constraints: top-view thresholding in the pre-filtering step and specification and calibration of a discrete choice model in the filtering step. In section 5 we show our results and we give concluding remarks in section 6.

2. SYSTEM OVERVIEW

In pedestrian tracking, there is a limitation in the detection systems where the target models can not be specified because background clutter may resemble foreground and background subtraction yields complex and noisy foreground regions, creating multiple peaks in a posterior distribution. However, many previous works minimize this kind of problems in multi-object tracking by placing the cameras at a high angle, looking down on the plane of motion of the objects (top-view plan).
In this context, we address the target detection problem by combining a behavioral model for pedestrian dynamic (see [4]), calibrated on real data, with a standard visual correlation technique between appearance models. An overview of our system is shown in fig 1. It is composed basically of three modules:

- **module 1** is focused on generating the hypotheses for the hypothetical moving object positions.
- **module 2** computes the well known correlation matching and for each successive frame the correspondence process attempts to associate the foreground regions with one of the existing tracks.
- **module 3** represents the principal contribution of this work : a behavioral-based filtering. Regions corresponding to real moving objects can be retained and false hypothesis can be rejected after a time period T with the pre-filtering and filtering steps.

We compute the visual displacement between a target region $\hat{r}^n_{t-1}$ at time $t-1$ and the associated region r^n_t at time t as the vector defined by the *maximum* of the correlation function between the two regions $C(\hat{r}^n_{t-1}, r^n_t)$. The visual displacement is then projected on the top-view plan using the parameters of the calibrated camera (height,camera axes inclination and focal distance). The series of top-view projections of visual displacements in successive frames gives rise to the pedestrian trajectory.
The update of the region of interest follows:

$$\hat{r}^n_t = \lambda \hat{r}^n_{t-1} + (1-\lambda) r^n_t \quad (1)$$

where the λ coefficient weights the contribution of the target and the associated regions. The hypothetical trajectories on top-view plan are then analyzed with the proposed behavioral model.

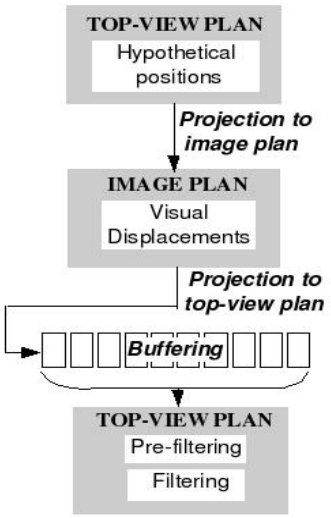

Fig. 1. Flow data in the proposed algorithm : filtering of hypothetical-moving objects on the top-view plan.

3. STARTING HYPOTHETICAL MOVING OBJECT POSITIONS

3.1. Background subtraction

In order to reduce background effects, the correlation is performed by using foreground image. Since the camera is fixed, the background can be modelled statistically. In order to detect the moving objects, we compute the difference between the background image and the current frame. Background subtraction will reduce the amount of data to be processed in motion estimation. From that, we obtain a binary support layer and a foreground-object image (see figure 2). We use the first one as a mask to place the hypothetical moving object positions and the second one to compute the visual displacements.

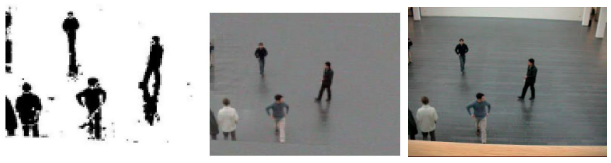

Fig. 2. **left**: binary mask, **center**: foreground, **right**: original frame.

3.2. Hypothetical moving object positions

We initialize the algorithm using a rectangular grid of points on the top-view plan with a resolution of 0.5 m. The grid is than projected back on the image plan and each hypothetical moving object position is filtered using the binary mask. This operation is illustrated in fig 4 (left). The choice of a top-view grid allows greater precision and keeps the possibility to fully use any a priori knowledge we can have on the scene. The algorithm projects the full grid only for the first frame as initialization step. After a period T (e.g. 10 frames), we repeat the procedure using a smaller grid placed along the border of the scene to be able to detect new incoming objects.

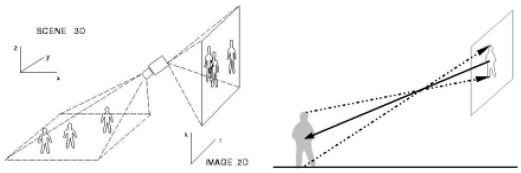

Fig. 3. **left**: the approximation of the Top-View plan by the image plan with a monocular camera, **right**: size estimation.

An alternative to the top-view grid could be a corner detector directly on the image. Experiments done using an Harris corner detector have shown that this technique is not flexible enough. Many corners, by definition, lie in the contour of the target object and they do not always correspond to a centroid in foreground regions. Moreover, with corners,it is not possible to have the control on the number of points.

3.3. Estimation of region size

For each hypothetical moving object position we have to associate a corresponding moving region (4, right). For simplicity, we use a rectangular region (*bounding box*). To compute the size of the bounding box, we suppose an averaged height of people equal to 160 cm and we ignore the error introduced by this approximation. In our practical application, we can estimate the size of the bounding box by projecting the hypothetical positions from top-view plan (see figure 3). This automatic scale selection is an useful tool to distinguish regions. In this way, for each visual tracker, we can perform a realistic partitioning (bounding boxes). We have used a calibrated camera so we know the camera parameters (height, camera axes inclination and focal distance).

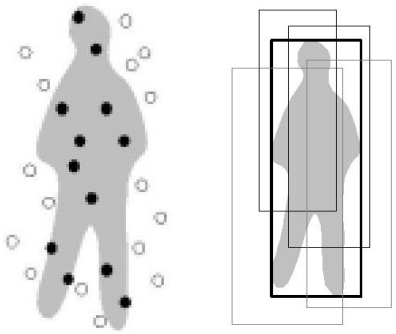

Fig. 4. **left**: Hypothetical positions on binary support layer, **right**: An example of moving regions.

4. BEHAVIORAL CONSTRAINTS

In this section we analyse the empirics and theory used to filter the motion trajectories. We split this task in two steps: *pre-filtering* and *filtering*.

4.1. Pre-filtering

In the *pre-filtering* process we estimate the trajectory state based on its past and current observations. The purpose of this step is to analize the hypothetical displacements on top-view plan with simple behavioral constraints. We think that a simple thresholding on the pedestrian speed and change in direction, as seen on the top-view plan, can be effective as compared to any other complex propagation model made directly on the image plan. So, in this preprocessing stage, we verify the projected displacements d_t^n and direction change $\Delta\theta_t^n$ of the hypothetical moving objects

$$d_t^n = \mathbf{p}_{\mathrm{t}}^{\mathrm{n}} - \mathbf{p}_{\mathrm{t-1}}^{\mathrm{n}}, \quad (2)$$

$$\Delta\theta_t^n = \theta_t^n - \theta_{t-1}^n \quad (3)$$

where $\mathbf{p}_{\mathrm{t}}^{\mathrm{n}}$ represents the position of visual tracker n at time t, and θ_t^n represents the direction of the displacement between the positions $\mathbf{p}_{\mathbf{t}}^{\mathbf{n}}$ and $\mathbf{p}_{\mathrm{t-1}}^{\mathrm{n}}$.
In this phase of the pedestrian tracking process, we keep a record for each visual tracker. The basic idea is to perform the pedestrian detection looking at pedestrian dynamic. In this spirit, we give a cumulative *score* to a pedestrian trajectory over an evaluation period T. We implement these ideas with simple thresholds on the projected displacement vectors defining:

$$I_t = \begin{cases} 0 & \text{if } \|d_t^n\| \leq \beta_d \text{ and } \|\Delta\theta_t^n\| \leq \beta_\theta \\ -1 & \text{otherwise} \end{cases}$$

where β_d and β_θ are the thresholds on one-step distance and direction change. The I_t is the one-step score given to a trajectory. We assign at each tracker a starting score equal to

S_0 and we decrement it at each 'bad' step. The tracker score S_T is evaluated over a period T assuming to accept a certain error ϵ on the trajectory (it means, practically, tollerate a certain number of 'bad' steps). We keep the tracker if the following condition is satisfied:

$$S_T = \frac{1}{T}\sum_{t=1}^{T} I_t \geq S_{inf} \quad (4)$$

where S_{inf} represents the minimum score for a good trajectory. In our experiments we use: $T = S_0 = 10$ and $\epsilon = \frac{S_0 - S_{inf}}{S_0} \geq 0.3$, that is equal to a margin of 30% (in this case we tollerate 3 'bad' steps over 10). Studies on pedestrian dynamics ([5]) show that the average speed value (in free-flow conditions) of a pedestrian is about 1.34 m/s. Our frame rate is 10 fps so we fix β_d to 13 cm. With analogous considerations we set β_θ to 120 degrees.

4.2. Filtering

The top-view thresholds on distance and change direction give us the input trajectories for the calibrated behavioral model.
We use discrete choise theory to build the behavioral model. This choice originates from the large success these techniques have found in other domains as market forecasting and traffic simulations. Discrete choice models in general and random utility models in particular are disaggregate behavioral models designed to forecast the behavior of individuals in choice situations.
In our case, a pedestrian is a decision maker involved in the choice process of *'where to put the next step'*. We consider that at each step, a pedestrian has a discrete space structure where he/she can walk on. The discretization of the space is *dynamic* and *individual-based*, that is, it depends on the current speed and direction of pedestrian (see fig 5). The origin point is the current position of pedestrian i while the distances *dmax*, *dmed* and *dmin* are related to the current individual speed V_i^c as follow:

$$V_i^{max} = V_i^c * 1.5; \quad dmax = V_i^{max} * \Delta T; \quad (5)$$
$$V_i^{med} = V_i^c; \quad dmed = V_i^{med} * \Delta T; \quad (6)$$
$$V_i^{min} = V_i^c * 0.5; \quad dmin = V_i^{min} * \Delta T; \quad (7)$$

ΔT is the time step between two successive observations in the data collection process for the model estimation.
Each cell represents an alternative which is described by an *utility* function. Such utility function is composed by a deterministic term, consisting of a combination of some *attributes* describing the alternatives, and a stochastic term that has to capture the correlation structure between the different alternatives (for more information about discrete choice models see [6],[7]). The output of the model is a set of probabilities describing each alternative. In this context we use the following expression for the utility function U_{ij} of the cell i, as perceived by decision maker j:

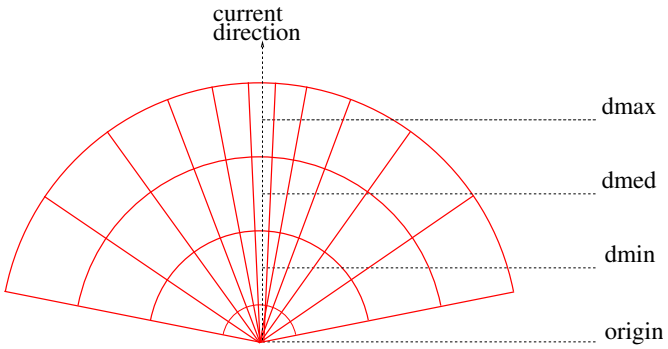

Fig. 5. The space model.

$$U_{ij} = \beta_1 * \textit{destination}_i + \beta_2 * \textit{direction}_i + \beta_3 * \textit{speed}_i \quad (8)$$

where the three attibutes are defined as follows:

1. *destination*: if we consider the triangle that has for vertex the current pedestrian position, the destination point (the last position in the current pedestrian trajectory) and the center of the cell i, the *destination* value is the angle at the current pedestrian vertex. It represents the angle between the cell i and the final destination.

2. *direction*: represents the angle between the cell i and the current pedestrian direction.

3. *speed*: is the module of the difference between the current pedestrian speed and the speed that characterizes the cell i.

The behavioral model has been calibrated using pedestrian trajectories manually grabbed from real video sequences and the values of β parameters estimated from this data. We have used a Cross Nested Logit model in such a way to have a more flexible structure in the correlation between alternatives (see [7]). In our model we have defined the nest structure considering that one alternative can be *accelerated*, *not accelerated*, *central* and *not central*. This correlation structure is illustrated in figure 6.
The model parameters have been estimated using the Biogeme package (see [8]). The general formulation of the CNL model is derived from the Generalized Extreme Value (GEV) model ([9]). The probability of choosing alternative i within the choice set C of a given choice maker is (see [7]):

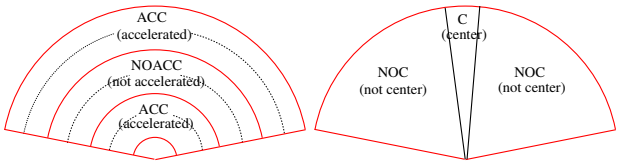

Fig. 6. **left:** Nesting based on speed, **right:** Nesting based on direction.

$$P(i|C) = \frac{y_i \frac{\partial G}{\partial y_i}(y_1, ..., y_J)}{\mu G(y_1, ..., y_J)} \quad (9)$$

basing on the following generating function:

$$G(y_1, ..., y_J) = \sum_m \left(\sum_{j \in C} \alpha_{jm} y_j^{\mu_m} \right)^{\frac{\mu}{\mu_m}} \quad (10)$$

The first derivative of G is:

$$\frac{\partial G}{\partial y_j} = \mu \sum_m \alpha_{jm} y_j^{\mu_m - 1} \left(\sum_{j \in C} \alpha_{jm} y_j^{\mu_m} \right)^{\frac{\mu}{\mu_m} - 1} \quad (11)$$

We fix the degrees of membership to the different nests (α_{jm}) to the constant value 0.5.
In our context, each step done by a pedestrian along its trajectory is characterized by a probability value. We give a mark to the trajectory k based on the cumulative value of probabilities and compare it with a threshold to filter out the *human trajectories* in the sense of our behavioral model:

$$M_k = \frac{\sum_{l=1, j \in C_n}^{l=L} P_{jl}}{\sum_{l=1, j \in C_n}^{l=L} \max P_{jl}} \geq th \quad (12)$$

where $j \in C_n$ is the alternative j in the choice set C_n, l is the current step and L is the number of steps in the trajectory k. The denominator is a normalization term so $0 \leq M_k \leq 1$.

5. RESULTS

We initialize the top-view grid on an area of $20 \times 100\ m^2$ at a resolution of 0.5 m and we work on a test sequence of 150 frames. We show the results of *pre-filtering* step in table 1 and those for *filtering* in figure 7.[1]
In the *filtering* step we fix the threshold to 0.2. This correspond to 30 trackers, correctly placed on pedestrians in the scene (see figure 9)[2]. In figure 8 we show an example of pre-filtered and filtered trajectories.

[1]The interested reader can find the elaborated video sequences at http://ltswww.epfl.ch/ltsftp/Venegas and http://ltswww.epfl.ch/ltsftp/antonini

Step	Number of trackers
Top-View grid $20 \times 100m^2$ at 0.5 m	8000
Foreground mask	598
Pre-filtering	68

Table 1. Results of *pre-filtering* step.

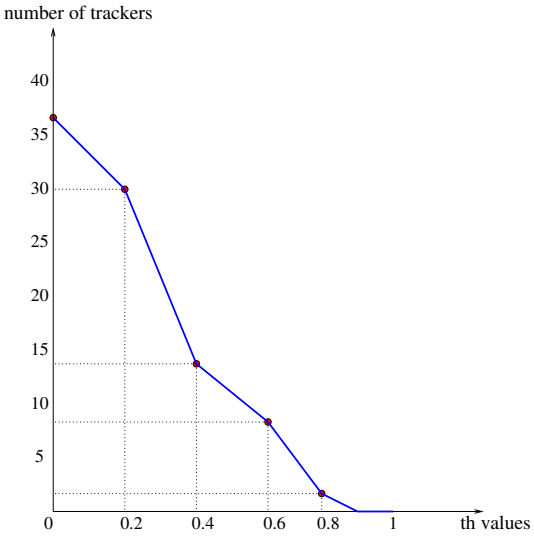

Fig. 7. The number of filtered trackers varying the threshold value of eq. 12.

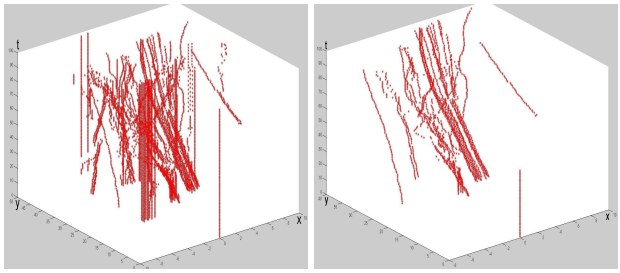

Fig. 8. **left**: Pre-filtered trajectories, **right**: Filtered trajectories.

The results show how the behavioral constraints allow the large number of starting hypothetical positions to converge towards the correct targets. We show in figure 10 the target regions associated with good and bad trajectories.

[2]To be precise, the bounding-boxes are displayed just when their movement fall in the pre-filtering limits. So the displayed trackers in fig.9 are less than 30, but the number of trajectories we keep is exactly the number resulting from the filtering step (30)

Fig. 9. A frame from the original sequence and one from the elaborated sequence.

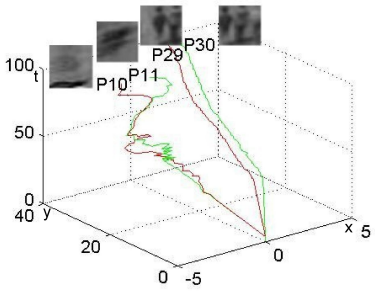

Fig. 10. Different target regions associated with different trajectories: trackers 10 and 11 are not well placed. On the contrary, targets 29 and 30 are well centered.

6. CONCLUSION AND FUTURE RESEARCH

In this paper we have approached the initiation step of the multi-target tracking problem using behavioral constraints formulated on the top-view plan and projected back on the image plan. This operation allows us to well detect pedestrians in cluttered environments. The approach is robust against bad image quality and resolution. A drawback of our approach is the presence of multiple trackers on the same target (for example, on different parts of the human body). We are currently working to merge the trajectories generated by these trackers to have a statistical evaluation about the number of pedestrians in the scene.
We aim also to improve the system by incorporating foreground modeling such as statistical models or shape-template. Finally, we aim to integrate an extended version of the behavioral model (taking into account interactions between individuals) in a non-linear non-gaussian state space framework. The purpose is to use the model to predict movements in the construction of human trajectories.

7. ACKNOWLEDGMENT

This work is supported by the Swiss National Science Fundation under the NCCR-IM2 project and by the Swiss CTI under project Nr. 6067.1 KTS, in collaboration with VisioWave SA, Ecublens, Switzerland.

References

[1] A. W. Senior. Tracking with probabilistic appearance models. In *ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems*, pages 48–55, 2002.

[2] M. Isard and A. Blake. Condensation: Unifying low-level and high-level tracking in a stochastic framework. In *European Conf. Computer Vision*, pages 893–908, 1998.

[3] G.Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[4] Michel Bierlaire, Gianluca Antonini, and Mats Weber. Behavioral dynamics for pedestrians. In K. Axhausen, editor, *Moving through nets: the physical and social dimensions of travel*, pages 1–18. Elsevier, 2003.

[5] M. Schreckenberg and S.D. Sharma, editors. *Pedestrian and Evacuation Dynamics*. Springer Verlag, 2002.

[6] Moshe Ben-Akiva and Michel Bierlaire. Discrete choice methods and their applications to short-term travel decisions. In Randolph Hall, editor, *Handbook of Transportation Science*, pages 5–34. Kluwer, 1999.

[7] Michel Bierlaire. A theoretical analysis of the cross-nested logit model. Accepted for publication in Annals of Operations Researchs, March 2001.

[8] Michel Bierlaire. An introduction to BIOGEME Version 0.6, February 2003.

[9] Daniel McFadden. Modelling the choice of residential location. In A.Karlquist, editor, *Spatial interaction theory and residential location*, pages 75–96, 1978.

[10] S.Venegas, S.F.Knebel, and J.P.Thiran. Multi-object tracking using particle filter algorithm on the top-view plan. *Submitted to ICASSP04*, 2003.