MULTI-CAMERA 3D PERSON TRACKING
WITH PARTICLE FILTER IN A SURVEILLANCE ENVIRONMENT

Jian Yao and Jean-Marc Odobez

IDIAP Research Institute
Centre du Parc, Av, des Pres-Beudin 20, 1920 Martigny, Switzerland
jyao@idiap.ch, odobez@idiap.ch, www.idiap.ch

ABSTRACT

In this work we present and evaluate a novel 3D approach to track single people in surveillance scenarios, using multiple cameras. The problem is formulated in a Bayesian filtering framework, and solved through sampling approximations (i.e. using a particle filter). Rather than relying on a 2D state to represent people, as is most commonly done, we directly exploit 3D knowledge by tracking people in the 3D world. A novel dynamical model is presented that accurately models the coupling between people orientation and motion direction. In addition, people are represented by three 3D elliptic cylinders which allow to introduce a spatial color layout useful to discriminate the tracked person from potential distractors. Thanks to the particle filter approach, integrating background subtraction and color observations from multiple cameras is straightforward. Alltogether, the approach is quite robust to occlusion and large variations in people appearence, even when using a single camera, as demonstrated by numerical performance evaluation on real and challenging data from an underground station.

1. INTRODUCTION

Person and object tracking is one of the most important tasks in a surveillance system. It allows to extract and link the semantic content of streams of video data, and is at the basis of the majority of higher level analysis algorithms. Despite tracking being one of the most studied topics in dynamic scene analysis, achieving good tracking performance in adverse conditions, as is often the case with indoor/metro surveillance settings, remains an important challenge.

In this paper we address the specific problem of single object tracking over a network of cameras with partial overlap. While multiple object tracking (MOT) is often desirable for automatic event analysis, tracking single person is also a task that is requested by end-users. More specifically, surveillance operators would like to be able to point at and 'tag' some particular person (or group of people) and track them through the metro station, so that there is always a monitor that displays a view of this person. This type of camera selection mechanism is very important for control rooms where the number of monitors is much less than the actual number of cameras in the network (e.g. 24 video monitors for more than 500 cameras in the Torino metro). This task has been coined the 'tag and track' task.

Several approaches have been proposed in the past for tracking people [1, 2, 3, 4, 5, 6]. Compared to 2D approaches [1, 2], methods relying on 3D state-spaces [3, 4, 5, 6] have been found to be more effective at modeling occlusion in the MOT case [3, 4, 6], and more robust in general, as they better constrain the solution space by allowing the introduction of more accurate priors (for instance, between the image position of the object and its size). However, until now, a large majority of research relying on 3D state-space (and 3D body representation) have been considered in closed indoor spaces (e.g. a lab room) [3, 4, 5], which do not present the same difficulties than surveillance videos (viewing angle is usually smaller, more overlap between cameras, etc). For instance, our attempt to apply [3] on our data failed mainly because obtaining good enough background subtraction was more challenging.

In [3, 7], 3D positions of humans are infered from multiple cameras. However, experiments are restrained to small spaces, as these methods rely more or less on volume intersection. In addition, many such approach have problems because they attempt at solving an inverse problem (reconstructing 3D from 2D images), and they can hardly be applied when using only two or even a single camera. In our case, thanks to the use of a particle filtering framework, the inverse problem is avoided, and the fusion of several cameras information is straightforward.

In this paper, we propose and evaluate a new method to address the 'tag & track' user scenario. The method relies on a particle filter approach with a 3D state representation. People are represented using 3 elliptic cylinders, allowing to introduce spatial layout information with respect to both background and more importantly color measurements, and which proved to be useful for distinguishing the tracked person from distracting people, for instance before and after occlusion. Through thorough numerical experiments on real and challenging data, we show that the approach is fairly robust, even when using a single camera.

The reminder of this article is organized as follows. We describe our method in Section 2. Section 3 presents our evaluation framework and numerical results. Section 4 concludes the paper.

2. 3D BODY MULTI-CAMER TRACKING WITH PARTICLE FILTER

In this Section, we introduce the single person 3D tracking algorithm based on a particle filter formulation. We will start by a quick presentation of the Bayesian filtering framework and its particle filter (PF) implementation, and then describe in more details the specific components that are involved in our implementation.

2.1 Bayesian tracking framework

The Bayesian formulation of the tracking problem is well known [8]. Denoting the hidden state representing the object configuration at time t by Y_t and the observation extracted from the image by Z_t, the objective is to estimate the filtering distribution $p(Y_t|Z_{1:t})$ of the state Y_t given the sequence of all the observations $Z_{1:t} = (Z_1, \dots, Z_t)$ up to the current time. Given standard assumptions, Bayesian tracking effectively solves the following recursive equation:

$$p(Y_t|Z_{1:t}) \propto p(Z_t|Y_t) \int_{Y_{t-1}} p(Y_t|Y_{t-1}) p(Y_{t-1}|Z_{1:t-1}) \mathrm{d}Y_{t-1} \qquad (1)$$

In non-Gaussian and non linear cases, this can be done recursively using sampling approaches, also known as particle filters (PF). The idea behind PF consists of representing the filtering distribution using a set of N_s weighted samples (particles) $\{Y_t^n, w_t^n, n = 1, ..., N_s\}$ and updating this representation when new data arrives. Given the particle set of the previous time step, $\{Y_{t-1}^n, w_{t-1}^n = \frac{1}{N_s}, n = 1, ..., N_s\}$, configurations of the current step are drawn from a proposal distribution $Y_t \sim q(Y|Y_{t-1}^n, Z_t)$. The weights are then computed as $w_t \propto w_{t-1}^n \frac{p(Z_t|Y_t)p(Y_t|Y_{t-1}^n)}{q(Y_t|Y_{t-1}^n, Z_t)}$. Finally, to avoid sampling im-

This work was supported by the European Union under the STREP project CARETAKER (http://www.ist-caretaker.org).

poverishment, it is necessary to apply an additional resampling step [8] whose effect is to eliminate the particles with low importance weights and to multiply particles having high weights.

Four elements are important in defining a PF:

- a state model which is an abstract representation of the object we are interested in;
- a dynamical model $p(Y_t|Y_{t-1})$ governing the temporal evolution of the state;
- a likelihood model $p(Z_t|Y_t)$ measuring the adequacy of the data given the proposed configuration of the tracked object;
- a proposal distribution $q(Y|Y^n_{t-1}, Z_t)$ the role of which is to propose new configurations in high likelihood regions of the state space. In our current case, we used a standard approach, using the dynamics as importance function, so that the the weight computation reduces to $w_t \propto w^n_{t-1} p(Z_t|Y_t)$.

These elements, along with our model will be described in the following Subsections.

2.2 3D State Space and body model:

The selection of an adequate state space is a compromise between two goals: on one hand, the state space should be precise enough so as to model as well as possible the information in the image and provide the richest information to further higher level analysis modules. On the other hand, it has to remain simple enough and in adequation to the quality level of the data (in our case, low to mid-level resolution) in order to obtain reliable estimates and keep the computation time low.

In the current situation, we decided to use a state space defined in the 3D space. This presents several advantages over a 2D approach. First, parameter setting, in most cases, will have a physical meaning. For instance, we can define a default height, the speed of an average walking person, etc. Besides, all prior information will automatically be 'built-in': for instance, according to the 3D position, we automatically know what should be the image size of the person image projection. Finally, occlusion reasoning -when tracking multiple people- would be simplified when using the 3D position.

In practice we modeled people using general cylinders, as illustrated in Fig. 1. Given the resolution of the images, we decided to use one cylinder for the head, one for the torso, and one for the legs. To account for the 'flatness'of people (or in other words, the width of people is usually larger than their thickness), we decided to use elliptic cylinders (i.e. the section of the cylinder is an ellipse). Utilizing this 3D human body model, one person standing on the ground plane with different orientation should produce different projected models in which the main difference is the width of the projected human bodies. Thus, in summary, the state space is represented by a 6-dimensional column vector:

$$X = [x, \dot{x}, y, \dot{y}, H, \alpha] \qquad \text{with} \tag{2}$$

- (x,y) denote the ground plane position of the object in the 3D physical space.
- $V = [\dot{x}, \dot{y}]^t$ denotes the speed of the object. The speed is characterized by its magnitude v expressed in cm.s^{-1}, and its direction γ^V (angle with respect to the X coordinate axis);
- H denotes the height of the object (in cm), and
- α denotes the orientation of the human body.

Figure 1 shows the body model along with the projection of the body model, for different state values, on one image.

2.3 The Dynamical Model

The dynamical model governs the temporal evolution of the state, and is defined as

$$\begin{aligned} p(X_k|X_{k-1}) &= p(x_k, \dot{x}_k|x_{k-1}, \dot{x}_{k-1}) p(y_k, \dot{y}_k|y_{k-1}, \dot{y}_{k-1}) && (3) \\ &\quad \times p(H_k|H_{k-1}) p(\alpha_k|\alpha_{k-1}, V_{k-1}) && (4) \end{aligned}$$

where we have assumed that the evolution of state parameters are independent given the previous state value. While this is reasonable

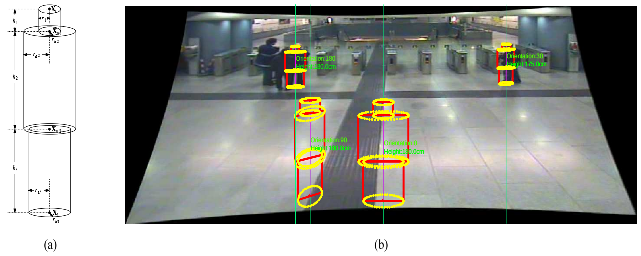


Figure 1: a) 3D body model consisting of three elliptic cylinders representing head, torso and legs. b) projection of the body model in the image (after distorsion removal) for different state values. Notice the change of width due to variation of the body orientation.

for height and orientation, the independence of the x and y variables is more questionable. In the future, we will investigate a coupling between these two variables. The specific models are the following. The position (and speed) along the x and y axis has been modeled as a Langevin motion, which corresponds to the motion of a free particle in a liquid with thermal excitation. The parameters of this model are most naturally specified in terms of continuous time parameters which have clear physical interpretation: the rate constant β, in s^{-1}, indicating the degree of friction in the liquid, and the stead-state root-mean square velocity $\bar{v}$ parameter, expressed in cm.s^{-1}. This model corresponds to the discrete dynamical model defined by

$$\dot{x}_k = a^x \dot{x}_{k-1} + b^x w^x_k, \qquad x_k = x_{k-1} + \tau \dot{x}_k \tag{5}$$

in which w^x_k are $\mathcal{N}(0,1)$ random variables, τ is the time step between two frames, and

$$a^x = \exp\left(-\beta^x \tau\right) \text{ and } b^x = \bar{v}^x \sqrt{1 - (a^x)^2} \tag{6}$$

By setting appropriate values for β and $\bar{v}$, using the above formula, parameter setting becomes independent of the actual frame rate (taken into account with τ).

The height model assumes that the height remains constant over time. In addition, it is constrained by a prior, to avoid large deviations towards too high or small values. The expression is:

$$p(H_k|H_{k-1}) \propto p_{temp}(H_k|H_{k-1}) p_{prior}(H_k) \tag{7}$$

where the first term imposes some temporal continuity to the estimated height, and is simply modeled as constant model (i.e. $p_{temp}(H_k|H_{k-1}) = \mathcal{N}(H_k; H_{k-1}, \sigma^2_{H_{temp}})$, where $\mathcal{N}(x; m, \sigma^2)$ denotes the value at x of a gaussian with mean m and variance σ^2, and $\sigma^2_{H_{temp}}$ denotes the noise variance of the temporal term) and the second term defines a prior over the height values, and is defined as $p_{prior}(H_k) = \mathcal{N}(H_k; H_0, \sigma^2_{H_0})$, where H_0 denotes an average reference height.

Finally, the body orientation dynamics is decomposed as:

$$\begin{aligned} p(\alpha_k|\alpha_{k-1}, V_{k-1}) &\propto p_{temp}(\alpha_k|\alpha_{k-1}) p_{motion}(\alpha_k|V_{k-1}) && (8) \\ p_{temp}(\alpha_k|\alpha_{k-1}) &= \mathcal{N}(\alpha_k; \alpha_{k-1}, \sigma^2_{\alpha_{temp}}) && (9) \\ p_{motion}(\alpha_k|V_{k-1}) &= \mathcal{N}(\alpha_k; \gamma^V_{k-1}, \sigma^2_{\alpha_{mot}}(v_{k-1})) && (10) \end{aligned}$$

Thus, on the one hand, the new orientation must account for the previous orientation value α_{k-1}, with a variance of $\sigma^2_{\alpha_{temp}}$. On the other hand, the second term p_{motion} constrains the orientation to align has much as possible with the direction of motion γ^V_{k-1}, and this constraint is controlled by the velocity v_{k-1}. Intuitively, when the person is moving with a high enough velocity, one would like the body orientation to be fully aligned with the direction of motion. At the other end, when a person is not moving, the direction of motion

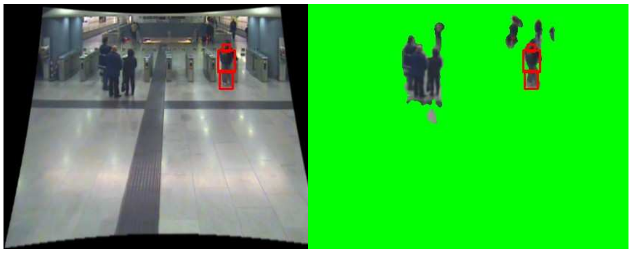


Figure 2: The 3 bounding boxes associated with a projection of a body model. Left: displayed on the original image (after distorsion removal). Right: displayed on the foreground image. To build the color histogram observations, only the visible foreground pixels are taken into account.

should play (almost) no role. This behaviour is achieved by selecting the variance $\sigma^2_{\alpha_{mot}}(v_{k-1})$ according to:

$$\frac{1}{\sigma^2_{\alpha_{mot}}(v_{k-1})} = \frac{1}{\sigma^2_{\alpha_{mot}}}\left(1+\frac{v_{k-1}}{v_{low}}\right) \tag{11}$$

where v_{low} denotes a small motion magnitude above which one wants to seriously align the body orientation with the direction of motion, and $\sigma^2_{\alpha_{mot}}$ is the default variance in absence of motion.

2.4 Observation Model

The observation model $\mathrm{p}(Z|X)$ measures the likelihood of the observation for a given state value, and is the main component of the tracker. In the following, we present its implementation in the single camera tracking case as well as its extension to multiple cameras.

Single camera case: The observations $Z = (Z^f, Z^{col})$ are composed of two parts: a foreground binary map Z^f obtained from background subtraction and color observations Z^{col} gathered in the form of multidimensional histograms. We have modeled the likelihood as the product of two terms:

$$\mathrm{p}(Z|X) = \mathrm{p}_{col}(Z^{col}|Z^f, X)\mathrm{p}_f(Z^f|X) \tag{12}$$

which we now describe. In both cases, the measurements rely on the bounding box $R_i, i = 1,2,3$ corresponding to each body part, as illustrated in Fig. 2.

Background component:

The foreground likelihood is modeled as:

$$\mathrm{p}_f(Z^f|X) = \mathrm{p}_f(\pi|X) = \prod_{b=1}^{3} \mathrm{p}(\pi_b) \tag{13}$$

where $\pi = [\pi_1, \pi_2, \pi_3]^t$, π_i denotes the percentage of pixels in R_i that belong to the foreground. The individual likelihood are modeled as:

$$\mathrm{p}(\pi_b) = \frac{1}{Cte}\exp\left(-\lambda_b^f A^b \varphi(\pi_b)\right) \tag{14}$$

where Cte is a normalization factor, and $\varphi()$ is defined as:

$$\varphi(x) = \begin{cases} 1 - T_{bot}, & x < T_{bot} \\ 1 - \exp(-\lambda_1 \cdot (T_{top} - x)), & x \leq T_{top} \\ 1 - \exp(-\lambda_2 \cdot (x - T_{top}), & x > T_{top} \end{cases} \tag{15}$$

where we define λ_1 and λ_2 values for different body parts. Alternatively, $\mathrm{p}(\pi_b)$ can be learned from training data. Notice that in Eq 14, the term in the exponential is weighted by the area A^b of the bounding box R_b. Thus, there will be less probability variations w.r.t. to observations for small regions than for large ones.

Color model :

As color models we used multidimensional color distributions represented by normalized histograms in the HSV space and gathered inside the candidate bounding boxes R_b associated with the state X_k. Note however that only the pixels labeled as foreground in Z^f are used to build the histograms. The use of different regions allows to add spatial layout information in the model. Consequently, the computation of the normalized multidimensional histogram results in a vector $\mathbf{b}(X_k) = (\mathrm{b}^j(X_k))_{j=1..N}$, where $N = 3 \times (N_h \times N_s \times N_V)$ with N_h, N_s and N_V representing the number of bins along the hue, saturation and value dimensions respectively ($N_h = N_s = N_V = 10$). At time t, the candidate color model $\mathbf{b}(X_t)$ is compared to a reference color model $\mathbf{b}_{ref}$. As a distance measure, we employed the Bhattacharyya distance measure [1, 2]:

$$D_{bhat}(\mathbf{b}(X_t), \mathbf{b}_{ref}) = \left(1 - \sum_{j=1}^{N}\sqrt{\mathrm{b}^j(X_t)\mathrm{b}^j_{ref}}\right)^{1/2} \tag{16}$$

and assumed that the probability distribution of the square of this distance for a given object follows an exponential law,

$$\mathrm{p}_{col}(Z^{col}|Z^f, X_k) \propto \exp\{-\lambda_{bhat} D^2_{bhat}(\mathbf{b}_k(X_k), \mathbf{b}_{ref})\}. \tag{17}$$

We used the histogram computed in the first frame as reference model [1, 2].

Multiple camera case: Due to the 3D state-space and particle filter approach, the only algorithmic modification to account for the availaibility of several cameras consists of modifying the likelihood term. We have now observations for each of the camera view: $Z = (Z_v)_{v=1...N_v}$, where each camera view observations are composed of foreground and color observations $Z_v = (Z_v^f, Z_v^{col})$, as in the single camera case, and N_v denotes the number of camera views where the object is visible. Assuming that the observation are independent given the state value, we can model the joint camera likelihood as:

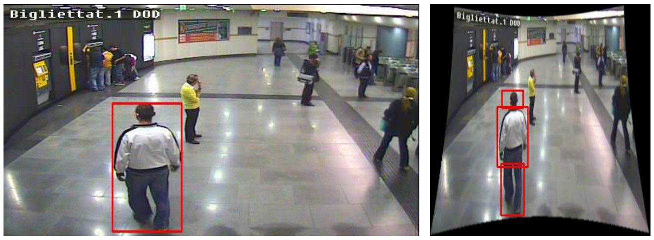
$$\mathrm{p}(Z|X) = \left(\prod_{v=1}^{N_v(X)} \mathrm{p}(Z_v|X)\right)^{\frac{1}{N_v(X)}} \tag{18}$$

where each $\mathrm{p}(Z_v|X)$ is itself defined according to Eq 12, and $N_v(X)$ denotes the number of camera in which the object located at state X is visible. Importantly, in this model, we have normalized the likelihood by taking the geometric mean of all the single view likelihoods, so that the overall likelihoods of states/objects which are visible in a different number of cameras are still comparable.

3. RESULTS

The evaluation protocol is the following:

Data We used three 2h30 minutes of video footage captured by three different cameras in the Torino metro station (one camera view is in Fig 2, Fig. 3 shows a second view. A third camera looks at the scene from a symetric position at an opposite location w.r.t. the camera of Fig. 3). The sequences are very challenging, due to the camera view points (small average people size and large people size variations in a given view, occlusion), and the presence of many



(a) (b)

Figure 3: Initialisation (a) manually marked bounding box on the original image, (b) the three bounding boxes from the 3D human body model corresponding to the initial state, projected on the undistorded image. Notice the strong specular reflections.

specular reflections on the ground which in combination with cast shadows generate many background subtraction false alarms. In addition most people are dressed with similar colors. The background subtraction was obtained using a robust multi-layer algorithm [9].
In order to compare the single and multiple camera tracking cases, we adopted the following annotation scheme. A set of ground truth tracks was annotated. For each track, the corresponding person was tagged from his entrance in the field of view of one camera until his disappearence of the same view. This segment of the ground truth will allow us to measure *on the same data* the improvement of the tracking due to the use of multiple cameras (E1 scenario). Then, the track ground truth was prolongated as long as the person remained visible in any of the 3 views until the person completely disappeared from the scene (E2 scenario). The annotation consisted of the person's bounding box and the camera number it appears in, gathered every second. A representative set of people were tagged in medium crowding situations, for a total of 36 people. We will denote E1 the single camera tracking scenario, where only the first part of the track is used (the person is visible in one given camera only), and E2 the multi-camera scenario where the full track is used as ground truth (the person is always visible *in the scene*). The average duration of one track is 35 seconds in the E1 scenario, and 46s when considering the full length of the tracks in scenario E2, for a total of around 8000 frames.

Performance measure: For each of the person, the bounding box sequence of the ground truth is compared with the bounding box sequence produced by the tracker. A coverage test is passed to assess whether the output region of the tracker matches that of the ground truth. It consists in computing the following measures:

$$\pi_{eval} = \frac{|B^{tr} \cap B^{gt}|}{|B^{tr}|}, \rho_{eval} = \frac{|B^{tr} \cap B^{gt}|}{|B^{gt}|}, \frac{1}{\mathrm{F}_{eval}} = \frac{1}{2}\left(\frac{1}{\rho_{eval}} + \frac{1}{\pi_{eval}}\right)$$

where $|.|$ denotes the set operator (the area), and π_{eval} and ρ_{eval} denote the precision and recall measures between the ground truth and tracker bounding boxes B^{gt} andB^{tr}, respectively. In order to have a good match, both these measures needs to be high, and this is reflected in the Fmeasure F_{eval}. Thus, if F_{eval} is higher than a threshold T_F, (in practice we used 0.2) we say that the tracker match the ground truth, otherwise not. Note that as soon as the coverage test is not satisfied at one instant, the tracking is said to have failed for the rest of the sequence. As performance measures, we report the percentage of times the tracker was able to successfully track the person up to T% of its full appearence duration in the camera (scenario E1) or in the scene (scenario E2).

The algorithm: Currently, assuming a 'tag-and-track' scenario, we assume that an initial bounding box is manually marked on the original image as shown in Figure 3(a). The bottom central point was used to initialize the person's location on the ground plane. The height of the transformed bounding box was used to compute initial person height. Based on this initial state, we project the used 3D human body model to get three bounding boxes projected onto the warped image as shown in Figure 3(b). These three projected bounding boxes are subsequently used to compute the object reference multi dimension histogram. **Importantly**, note that for the multi-camera case, only this single reference histogram computed from the initialization view is used, even if the person is seen from another camera view in the initial state. Finally, all model parameters were set to standard values and kept identical for all experiments.

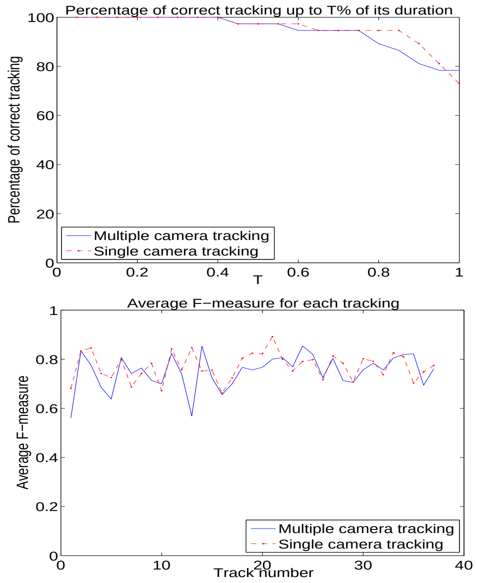


Figure 4: E1 scenario. Comparison of the tracking results when using one or multiple cameras. Top) the tracking success rates. Bottom) for each track, the average F measure F_{eval}, computed only on the frames until the failure occurs.

Results: Figure 4a) presents the tracking results in the E1 scenario. Overall, the results are quite good, demonstrating the robustness of the algorithm, even with a single camera, and despite the presence of full occlusions in the majority of the tracks. Surprisingly, the use of multiple cameras is performing similarly to the single camera case, although the errors do not appear for the same tracks. While the use of multiple cameras allows to correctly track several people that were missed when using a single camera, it also introduces some rare and spurious errors when the number of views in which the tracked person appears is changing (esp. increasing), i.e. the person is about to appear in a new view. In this situation, the algorithm usually gives preference to particles with less measurements, as they may provide a better (geometric) average likelihood (cf Eq. 18) since the current 3D state estimate might not allow for good person localization in the new view[1], and the tracker is momentarily 'locked' to match a single view mesurements, and the track is lost. To address this issue, one approach could be to enforce having all particles at each time step to rely on the same number of view measurements, or to exploit a proposal that would include information from the new view. Note however that multiple cameras allows to cover more space and track people for longer periods, and that from other qualitative experiments we conducted, it was shown that multi-camera improves the results and the robustness of the tracking (e.g. we were able to correctly track a person for more than 10 minutes in the hall in presence of many simultaneous occlusions, which was not possible with a single camera). Also, while 2D localization was not improved in the 2D view that was ground truthed to evaluate the precision of the tracking (Fig 4b)), we could observed a better 3D localization on the ground plane (which is more difficult to evaluate though).
To study the compromise between computational ressources and tracking robustness, we evaluated the results by varying the number of particles in the sampling approximation. As expected, there is a drop in performance when diminishing this number, but it is not dramatic. Also, when using more than 200 hundred particles, we can see that there is no specific improvements (differences can be due to random fluctuations since the tracking process is stochastic). With 100 particles, the algorithm runs at approximately 4 to 2 frame per second for single camera tracking, and around 1 to 2 frame/sec in the multi-camera case. Most of the time is indeed spent in the background subtraction algorithm. Finally, Figure 6 shows than when tracking the people in the whole scene, the performance remains approximately the same. One can notice a performance drop at the end of the tracks. This is mainly due to the fact that most of the tracks end up beyond the gates (see camera view in Fig. 2): the person is therefore very small, partially occluded (bottom of the legs not visible), and visible only in one view.

Finally, Figure 7 illustrates on one typical example the tracking result with a single camera.

[1] Note also that the color model was not learned on the new view as well.

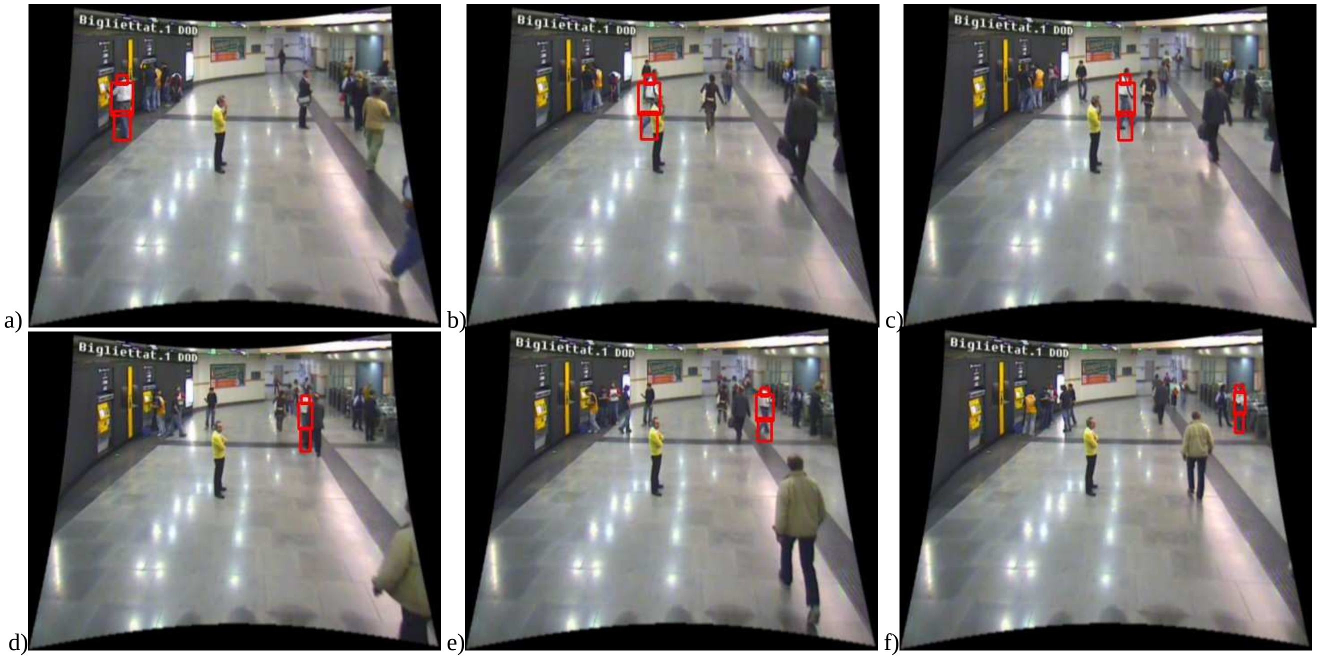
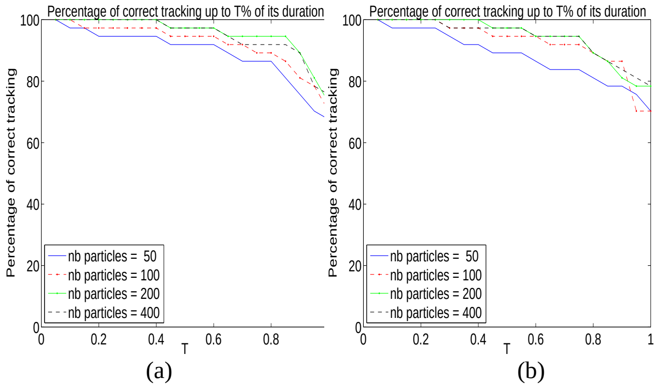


Figure 7: Single camera tracking result. Initialization at time 0s in Fig. 3. results, at time 4.6s, 28s, 29s, 31s, 32s, 34s. a) the person is buying tickets at the vending machine, b) leaving the vending machine, partial occlusion d) and e) before and after full occlusion.



(a) (b)

Figure 5: E1 scenario. Tracking performance in function of the number of particles used. a) single camera tracking b) multiple camera tracking.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed and evaluated a new method to address the 'tag & track' user scenario. The method relies on a particle filter approach with a 3D state representation and 3D people modeling. Numerical experimentation on real challenging surveillance data showed that our approach is quite robust, even with a single camera.

Several issues remain to be investigated. One is the development of a simple interface to initialize the tracking from a live video (e.g. by a single click on a person head). Accordingly, we need to investigate the influence of the approximate initialisation on the tracking performance. Also, to improve the tracking, we would like to study the use of several color models (since they may change depending on the camera view) as well as their adaptation (e.g. during the instant when the person is isolated in a view).

REFERENCES

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2000, pp. II:142–149.

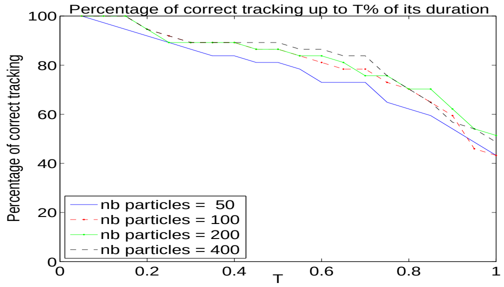


Figure 6: E2 scenario. Tracking performance in function of the number of particles used, for multiple camera tracking.

[2] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. of* 7^{th} *Eur. Conf. Comp. Vision*, Denmark, June 2002, pp. 661–675.

[3] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *accepted for publication in IEEE PAMI*, 2007.

[4] T. Osawa, X. Wu, K. Wakabayashi, and T. Yasuno, "Human tracking by particle filtering using full 3d model of both target and environment," in *proc. of ICPR*, 2006.

[5] O. Lanz, "Approximate bayesian multibody tracking," *IEEE PAMI*, vol. 28, no. 9, pp. 1436–1449, Sept. 2006.

[6] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment," in *Proc. of CVPR*, Washington DC, June 2004.

[7] A. Mittal and L. S. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," in *ECCV*, 2002.

[8] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[9] J. Yao and J-M. Odobez, "Multi-layer background subtraction based on color and texture," in *Proc. IEEE CVPR Workshop on Visual Surveillance (CVPR-VS)*, June 2007, pp. 1–8.