Multi-Camera Multi-Person 3D Space Tracking with MCMC in Surveillance Scenarios[⋆]

Jian Yao Jean-Marc Odobez

Idiap Research Institute
Centre du Parc, Rue Marconi 19, CH-1920 Martigny, Switzerland
E-Mails: {Jian.Yao,Jean-Marc.Odobez}@idiap.ch

Abstract. We present an algorithm for the tracking of a variable number of 3D persons in a multi-camera setting with partial field-of-view overlap. The multi-object tracking problem is posed in a Bayesian framework and relies on a joint multi-object state space with individual object states defined in the 3D world. The Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) method is used to efficiently search the state-space and recursively estimate the multi-object configuration. The paper presents several contributions: i) the use and extension of several key features for efficient and reliable tracking (e.g. the use of the MCMC framework for multiple camera multiple object tracking; the use of powerful human detector outputs in the MCMC proposals to automatically initialize/update object tracks); ii) the definition of appropriate prior on the object state, to take into account the effects of 2D image measurement uncertainties on the 3D object state estimation due to depth effects; iii) a simple rectification method aligning people 3D standing direction with 2D image vertical axis, allowing to obtain better object measurements relying on rectangular boxes and integral images; iv) representing objects with multiple reference color histograms, to account for variability in color measurement due to changes in pose, lighting, and importantly multiple camera view points. Experimental results on challenging real-world tracking sequences and situations demonstrate the efficiency of our approach.

1 Introduction

Multiple object tracking (MOT) in video is one of the fundamental research topics in dynamic scene analysis, as tracking is usually the first step before applying higher level scene analysis algorithms. While fairly good solutions to the tracking of isolated objects or small number of objects having transient occlusion have been proposed in the past, MOT remains challenging with higher densities of people, mainly due to inter-person occlusion, bad observation viewpoints, small resolution images, entering/leaving of people, etc. These situations are often encountered in the visual surveillance domain.

There is an abundance of literature devoted to MOT. In past years, state-space models [1–3] have been shown to be the most successful. Although some methods choose to use a single-object state-space model [3], only a more rigorous formulation of the MOT problem using a joint state space model allows object interactions and identity to be properly defined. In general, interactions is defined based on proximity, occlusion being so far the most studied problem. Tracking a variable number of objects with particle filters (PF) has been addressed in [1, 4–6]. These works highlighted the need for

[⋆] This work was supported by the European Union 6th FWP Information Society Technologies CARETAKER project (Content Analysis and Retrieval Technologies Applied to Knowledge Extraction of Massive Recordings, FP6-027231).

a global observation model to deal with multi-object configurations varying in number, in order to obtain likelihoods of the same order of magnitude for configuration with different number of objects.

To alleviate the occlusion problem in medium to crowded scenes, the use of multiple cameras and the fusion of the information between them becomes almost necessary [7–9]. Fleuret *et al.* [9] proposed an algorithm that can reliably track multiple persons in a complex environment and provide metrically accurate position estimates by combining a probabilistic occupancy map. Du and Piater [8] present a novel approach to ground-plane tracking of targets in multiple cameras by using collaborative particle filters. This method performs very well and can handle the imprecise foot positions and some calibration uncertainties. Regretfully, the current approach is available only for single object tracking.

In this paper, we propose a novel algorithm to automatically detect and track a variable number of people in a multi-camera environment with partial field of view overlap. More precisely, we adopt a multi-object state space Bayesian formulation, solved through RJ-MCMC sampling for efficiency reasons [4, 6]. The proposal (i.e function sampling new state configurations to be tested) used in this scheme takes advantage of a powerful machine learning human detector allowing to efficiently update tracks or initialize new tracks. We adopt a 3D approach where object states are defined in a common 3D space allowing to represent people with a body model, and to facilitate occlusion reasoning. The multi-camera fusion is solved by using global likelihood models over foreground and color observations. Our contributions are to combine and integrate efficient algorithmic components in our framework which have been shown in the past (often separately) to be essential for accurate and efficient tracking, and to propose additional techniques to solve specific issues as detailed below.

One issue in multiple object tracking is object interaction modeling. This can be done by defining priors over the joint state space. Such prior is usually based on object proximity, which prevents objects of occupying the same state space region or explaining twice the same piece of data. In our case, we propose to refine such priors by exploiting both the body orientation in the definition of proximity, and by using the prediction of the future object state to model that moving people tend to avoid colliding each other.

Multi-camera tracking in surveillance scenarios is usually quite different than tracking in indoor rooms. Larger field-of-view (FOV) cameras are used to cover more physical space, the overlaps between the FOV are smaller, and people appear with dramatically different image resolutions due to their placements and points of views. As a consequence, a small and seemingly not significative 2D position change (e.g. one pixel) in one view can correspond to a large position change in the other view, as illustrated in Fig. 2. This is particularly problematic at *transitions* between FOV cameras, when a person enters a new view which has much higher resolution than the current one. As due to this uncertainty, the projection of the current estimate does not match at all the person in the new view. As a result, the tracker will assume that the person remains only in the first view, and will initialize a new track in the new view. To solve this issue, we propose to integrate in the 3D object state prior a component which models the effects of the image estimation uncertainties according to the views in which the object is visible, and to use a proposal taking into account the human detection output per view to draw samples at well localized places in the new view.

One final novelty of the paper is an image rectification step allowing to reduce people geometric appearance variability in images due to the use of of large FOV cameras.

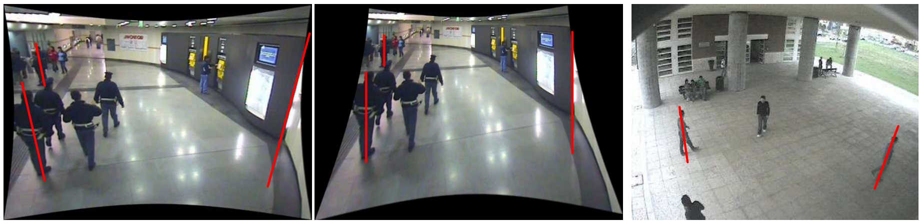


Figure 1: *Vertical vanishing point mapping. Left: after distorsion removal and before the mapping. We can observe people slant according to their position. Central: after the mapping to the infinity. Bounding-boxes will fit more closely the silhouette of people. Right: another example.*

More precisely, in these cases, people often appear slanted in the border of an image, even after the removal of radial distorsions, as illustrated in Fig. 1. This is a problem for human detectors which often consist of applying a classifier on rectangular regions, or in other tasks (e.g. tracking) when integral images are used to efficiently extract features over boxes, as the variation of people orientation in the image will affect the consistency of the extracted features (with respect to an upright standing) and will ultimately harm the detection or tracking processes. To remove this variability, we propose a simple rectification scheme which is applied to the input image as a pre-processing step. It consists of mapping the 3D vertical lines into 2D vertical image lines, as illustrated in Fig. 1. The method is shown to introduce negligible image distorsions, and can be applied in any scenarios where an initial calibration step is feasible.

The rest of this paper is organized as follows. Section 2 presents our slant removal rectification procedure. The multi-camera multi-person tracking framework is described in Section 3, along with its main features. Experimental results on real data are reported in Section 4, followed by the conclusion in Section 5.

2 Calibration and Vertical Vanishing Point Mapping

Camera Calibration: Cameras were calibrated using the available information and exploiting geometrical constraints [10], like 3D lines should appear as undistorted, or vertical direction Z is obtained from the image coordinates of the vertical vanishing point $\mathbf{v}_\perp$, computed as the intersection of the image projections of a set of 3D world parallel vertical lines. The image-to-ground homography $\mathbf{H}$ was estimated using a set of manually marked points in the image plane and their 3D correspondences in the 3D ground plane.

Removing Slant by Mapping the Vertical Vanishing Point to Infinity: In Fig. 1, we observe that standing people appear with different slants in the image. This introduces variability in the feature extraction process when using rectangular regions. To handle this issue, we propose to use an appropriate projective transformation $\mathbf{H}_\perp$ of the image plane in order to map its vertical finite vanishing point to a point at infinity. As a result, the 3D vertical direction of persons standing on the ground plane will always map to 2D vertical lines in the new image, as illustrated in Fig. 1. This transformation should thus help in obtaining better detection results or extracting more accurate features while still keeping the computation efficiency, e.g. by using integral images.

Our goal is to find a 2D homography $\mathbf{H}_\perp$ that maps the image vertical vanishing point $\mathbf{v}_\perp = (x_\perp, y_\perp, 1)^\top$ to a vanishing point at infinity $(0, y_\infty, 0)^\top$ where y_∞ can be any non-zero value. As the above mapping alone is not sufficient to fully define the homography, we must enforce additional constraints in order to avoid severe projective distortions of the image. To obtain a resampled image that looks as much as possible like the original image, we enforce that the transformation $\mathbf{H}_\perp$ should act as far as possible as a rigid transformation in the neighborhood of a given selected point $\mathbf{x}_0$ of the image. This means that the first order approximation of the transform in the neighborhood of $\mathbf{x}_0$ should be a rotation rather than a general affine transform. An appropriate choice of $\mathbf{x}_0$ to enforce such as constraint can be the image center.

For the moment, assume that $\mathbf{x}_0$ is the coordinate system origin, and that $\mathbf{v}_\perp$ is located on the y axis, i.e. $\mathbf{v}_\perp = (0, y_\infty, 1)^\top$. Then we can consider the homography $\mathbf{G}$ which maps the vertical vanishing point to a point at infinity $(0, y_\infty, 0)^\top$ as required, and maps a 2D point (x, y) into a 2D point (x', y') according to:

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{-1}{y_\infty} & 1 \end{bmatrix}, \begin{bmatrix} x' \\ y' \end{bmatrix} = \frac{y_\infty}{(y_\infty - y)} \begin{bmatrix} x \\ y \end{bmatrix}, \frac{\partial(x', y')}{\partial(x, y)} = \frac{y_\infty}{(y - y_\infty)^2} \begin{bmatrix} y_\infty - y & x \\ 0 & y_\infty \end{bmatrix}.$$

The last part above provides the Jacobian of the transform and models the linear distorsions. It shows that, at the origin $(0, 0)$, the Jacobian is equal to the identity matrix, which means that no linear distorsions are introduced by the transform at this point.

In the general case, it is easy to show that for an arbitrarily placed point of interest $\mathbf{x}_0 = (x_0, y_0, 1)^\top$ and vertical vanishing point $\mathbf{v}_\perp = (x_\perp, y_\perp, 1)^\top$, we can reach the above special case by applying first the translation $\mathbf{T}$ that maps the origin of the coordinate system to the selected point $\mathbf{x}_0$, and then the rotation $\mathbf{R}$ which brings the translated vertical vanishing point on the y axis. The required mapping $\mathbf{H}_\perp$ is then given by $\mathbf{H}_\perp = \mathbf{GRT}$, and can be used to warp the undistorted image and obtain the wanted image (central image in Fig. 1). Accordingly, the new image-to-ground homography $\hat{\mathbf{H}}$ can be updated (e.g. $\hat{\mathbf{H}} = \mathbf{H}\mathbf{H}_\perp^{-1}$).

3 Multi-Camera Multi-Person Tracking

In this section, we introduce the multi-camera multi-person 3D tracking algorithm based on a Markov Chain Monte Carlo (MCMC) sampling method, and then provide the main elements of the model, focusing on the main aspects of our approach.

3.1 Bayesian Tracking Framework and 3D Multi-Person State Representation

In the Bayesian framework, the goal is to estimate the conditional probability $p(\tilde{\mathbf{X}}_t|\mathbf{Z}_{1:t})$ of the joint multi-person configuration $\tilde{\mathbf{X}}_t$ at time t given the sequence of observations $\mathbf{Z}_{1:t} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_t)$. This posterior probability $p(\tilde{\mathbf{X}}_t|\mathbf{Z}_{1:t})$, known as the filtering distribution, can be expressed recursively using the Bayes filter equation:

$$p(\tilde{\mathbf{X}}_t|\mathbf{Z}_{1:t}) = \frac{1}{C} p(\mathbf{Z}_t|\tilde{\mathbf{X}}_t) \times \int_{\tilde{\mathbf{X}}_{t-1}} p(\tilde{\mathbf{X}}_t|\tilde{\mathbf{X}}_{t-1}) p(\tilde{\mathbf{X}}_{t-1}|\mathbf{Z}_{1:t-1}) d\tilde{\mathbf{X}}_{t-1}, \tag{1}$$

where the dynamical model, $p(\tilde{\mathbf{X}}_t|\tilde{\mathbf{X}}_{t-1})$, governs the temporal evolution of the joint state $\tilde{\mathbf{X}}_t$ and the observation likelihood model $p(\mathbf{Z}_t|\tilde{\mathbf{X}}_t)$ measures the fitting accuracy of the observation data $\mathbf{Z}_t$ given the joint state $\tilde{\mathbf{X}}_t$. C is a normalization constant. In non-Gaussian and non-linear cases, the filter equation can be approximated using Monte Carlo methods, in which the posterior $p(\tilde{\mathbf{X}}_t|\mathbf{Z}_{1:t})$ is represented by a set of N samples $\{\tilde{\mathbf{X}}_t^{(r)}\}_{r=1}^N$. For efficiency, in this work we use the Markov Chain Monte Carlo (MCMC) method, where the set of samples have equal weights and form a so-called Markov chain. Consequently, we obtain the following Monte Carlo approximation:

$$p(\tilde{\mathbf{X}}_t|\mathbf{Z}_{1:t}) \approx \frac{1}{C} p(\mathbf{Z}_t|\tilde{\mathbf{X}}_t) \times \sum\nolimits_{r=1}^{N} p(\tilde{\mathbf{X}}_t|\tilde{\mathbf{X}}_{t-1}^{(r)}). \tag{2}$$

Single Object 3D State and Model: We modeled people using general cylinders, as illustrated in Fig. 3. Given the resolution of the images, we decided to use three cylinders: one for the head, one for the torso, and one for the legs. We used elliptic cylinders (i.e. the section of the cylinder is an ellipse) to account for people 'flatness' (people width is usually larger than their thickness), which allows to produce different image projected models depending on people's orientations w.r.t. the camera. The state space of a human person i at time t is represented by a 6-dimensional vector:

$$\mathbf{X}_{i,t} = \left(x_{i,t}, y_{i,t}, \dot{x}_{i,t}, \dot{y}_{i,t}, h_{i,t}, \alpha_{i,t}\right)^\top, \tag{3}$$

where $\mathbf{u}_{i,t} = (x_{i,t}, y_{i,t})^\top$ denotes the person ground plane position in the 3D physical space. $\dot{\mathbf{u}}_{i,t} = (\dot{x}_{i,t}, \dot{y}_{i,t})^\top$, h and $\alpha_{i,t}$ denote the speed, the height, and the orientation w.r.t. the X-direction on the ground plane, respectively.

The Multi-Object State Space is defined as:

$$\tilde{\mathbf{X}}_t = (\mathbf{X}_t, \mathbf{k}_t), \tag{4}$$

where $\mathbf{X}_t = \{\mathbf{X}_{i,t}\}_{i=1\dots M}$, M is the maximum number of objects appearing in the scene at any given time instant, and $\mathbf{k}_t = \{k_{i,t}\}_{i=1\dots M}$ is a M-dimensional binary vector. The boolean value $k_{i,t}$ signals whether the object i is valid/exists in the scene at time t ($k_{i,t} = 1$), or not ($k_{i,t} = 0$). The identifier set of existing objects is thus represented as $\mathcal{K}_t = \{i \in [1, M] | k_{i,t} = 1\}$, and $\bar{\mathcal{K}}_t = \{1, 2, 3, \dots, M\} \setminus \mathcal{K}_t$.

3.2 Dynamical Model

The dynamical model governs the evolution of the state between time steps. It is responsible for predicting the motion of people as well as modeling inter-personal interactions between the various people.

The Joint Dynamical Model for a variable number of people is defined as follows:

$$p(\tilde{\mathbf{X}}_t|\tilde{\mathbf{X}}_{t-1}) \propto p_0(\mathbf{X}_t|\mathbf{k}_t) \left(\prod\nolimits_{i=1}^{M} p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1}, \mathbf{k}_t, \mathbf{k}_{t-1}) \right) p(\mathbf{k}_t|\mathbf{k}_{t-1}, \mathbf{X}_{t-1}) \tag{5}$$

$$\text{with } p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1}, \mathbf{k}_t, \mathbf{k}_{t-1}) = \begin{cases} p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}) & \text{if } \; i \in \mathcal{K}_t \text{ and } i \in \mathcal{K}_{t-1}, \\ p_{birth}(\mathbf{X}_{i,t}) & \text{if } \; i \in \mathcal{K}_t \text{ and } i \notin \mathcal{K}_{t-1} \text{(birth)}, \\ p_{death}(\mathbf{X}_{i,t}) & \text{if } \; i \notin \mathcal{K}_t \text{ and } i \in \mathcal{K}_{t-1} \text{death)}. \end{cases}$$

The term $p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1})$ denotes a single person dynamics, as discussed later, while p_{birth} and p_{death} denote prior distributions over the state space for newborn or dead objects. The last term $p(\mathbf{k}_t|\mathbf{k}_{t-1}, \mathbf{X}_{t-1})$ in Eq. (5) allows to define a prior over the number of objects which die and are born at a given time step, thus disfavoring for instance the deletion of an object and its replacement by a newly created object.

Shape oriented and person avoidance interactions prior. Person interactions are modeled by the the term p_0 in Eq. (5) and defined by pairwise prior over the *joint* state space:

$$p_0(\mathbf{X}_t|\mathbf{k}_t) = \prod\nolimits_{i,j \in \mathcal{K}_t, i \neq j} \phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) \propto \exp\left\{ -\lambda_g \sum\nolimits_{i,j \in \mathcal{K}_t, i \neq j} g(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) \right\},$$

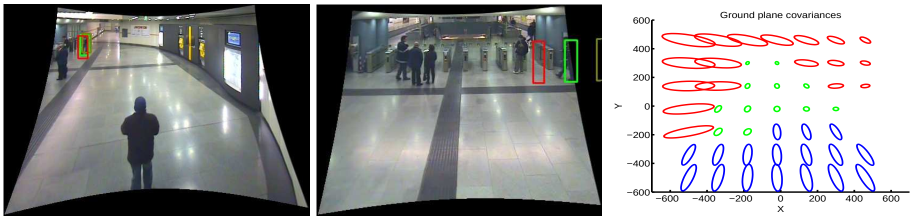


Figure 2: *Left images. Due to depth effects, very similar positions in the first camera view corresponds to dramatically different image locations on the other view. Right graph. for the same scene, floor map of localization uncertainties, propagated from image localization uncertainties. In green, floor locations visible in both cameras. In blue/red, locations visible by only one camera.*

where $g(\mathbf{X}_{i,t}, \mathbf{X}_{j,t})$ is a penalty function. In papers [4, 6] which used such a prior, authors defined this penalty function based on the current 2D overlap between the object projections, or on the euclidian distance between the two object centers, for instance, $g(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) = \psi(\|\mathbf{u}_{i,t} - \mathbf{u}_{j,t}\|)$. In our case, we propose two improvements: first, as people are not 'circular', we replaced the above euclidian distance by a Mahalanobis distance, i.e. $g_p(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) = \psi\left(d_{m,i}(\mathbf{u}_{i,t} - \mathbf{u}_{j,t})\right) + \psi\left(d_{m,j}(\mathbf{u}_{i,t} - \mathbf{u}_{j,t})\right)$ where $d_{m,i}$ (resp. $d_{m,j}$) is the Mahalanobis distance defined by the ellipsoid shape of the person i (resp. j). Qualitatively, this term favors the alignment of people orientation (of close by people) in contrast to people with perpendicular orientations. People following each other is a typical situation where this term can be useful.

Secondly, when people move, they usually look forward to *avoid collision* with other people. We thus introduced a prior as well on the state $\mathbf{X}^{pr}_{i,t+1}$ predicted from the current state value $\mathbf{X}_{i,t}$, by defining the penalty function as $g(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) = g_p(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) + g_p(\mathbf{X}^{pr}_{i,t+1}, \mathbf{X}^{pr}_{j,t+1})$. This term will thus prevent collision, not only when people are coming close, but also when people are moving together in the same direction.

The Dynamical Model of a Single Person is defined as:

$$p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}) = p(\mathbf{u}_{i,t}, \dot{\mathbf{u}}_{i,t}|\mathbf{u}_{i,t-1}, \dot{\mathbf{u}}_{i,t-1}) p(h_{i,t}|h_{i,t-1}) p(\alpha_{i,t}|\alpha_{i,t-1}, \dot{\mathbf{u}}_{i,t}), \quad (6)$$

where we have assumed that the evolution of state parameters is independent given the previous state values. In this equation, the height prior $p(h_{i,t}|h_{i,t-1})$ assumes a constant height model with a steady-state value, to avoid large deviations towards too high or small values. The body orientation dynamics $p(\alpha_{i,t}|\alpha_{i,t-1}, \dot{\mathbf{u}}_{i,t})$ is composed of two terms which favor temporal smoothness and orientation alignment with the walking speed (which depends on the speed magnitude) as we have described in the single person tracking algorithm [11].

In addition to prior terms which prevent invalid floor positions for people and reduce the likelihood of the state when the walking speed exceeds some predefined limit, the position/speed dynamics is defined by

$$\dot{\mathbf{u}}_t = \mathbf{A}\dot{\mathbf{u}}_{t-1} + \mathbf{B}\mathbf{w}_{1,t} \quad \text{and} \quad \mathbf{u}_t = \mathbf{u}_{t-1} + \tau\dot{\mathbf{u}}_t + \mathbf{C}(\mathbf{u}_{t-1})\mathbf{w}_{2,t} \quad (7)$$

where $\mathbf{w}_{q,t} = (w^{(x)}_{q,t}, w^{(y)}_{q,t})^\top$ is a Gaussian white noise random variable ($q = 1, 2$), and τ is the time step between two frames. First assume that $\mathbf{C}(\mathbf{u}_{t-1}) = 0$. In this case, Eq. (7) models a Langevin motion, with $\mathbf{A} = a\mathbf{I}$ and $\mathbf{B} = b\mathbf{I}$ ($\mathbf{I}$ denotes the 2×2 identity matrix) and $a = \exp(-\beta\tau)$ and $b = \bar{v}\sqrt{1 - a^2}$, where β accounts for the speed damping and $\bar{v}$ is the steady-state root-mean square speed.

Introducing 2D-to-3D localization uncertainties. In multi-view environments with small overlapping regions between views, and important depth scene effects with large image projection size variations of people within and across views (see Fig. 2), the Langevin motion is not enough to represent the state dynamics uncertainty. Fig. 2 illustrates a typical problem at view transitions: a person appearing at a small scale in a given view enters a second view. Observations from the first view are insufficient to accurately localize the person on the 3D ground plane. Thus, when the person enters the second view, the image projections obtained from the state prediction of the MCMC samples will often results in a mismatch with the actual localization of the person in the second view. This mismatch might be too high to be covered (in one time step) by the regular noise of the dynamical model. As a result, the algorithm may keep (for some time) the person track so that it is only visible in the first view, and create a second track to account for the person's presence in the second view. To solve this issue, we added the noise term $\mathbf{C}(\mathbf{u}_{t-1})\mathbf{w}_{2,t}$ on the location dynamics, whose covariance magnitude and shape depend on the person location. The covariance of this noise, which models 2D-to-3D localization uncertainties, is obtained as follows. The assumed 2D Gaussian noises on the image localization of a person's feet from the different views are propagated to the 3D floor position using an Unscented Transform, and potentially merged for people positions visible from several cameras, leading to the pre-computed noise model illustrated in the right image of Fig.2. Qualitatively, this term guarantees that in the MCMC process, state samples drawn from the dynamics will actually spread the known uncertainty 3D regions, and those samples drawn by exploiting the human detectors will not be disregarded as being too unlikely according to the dynamics.

3.3 Observation Model

When modeling $p(\mathbf{Z}_t|\tilde{\mathbf{X}}_t)$, which measures the likelihood of the observation $\mathbf{Z}_t$ for a given multi-object state configuration $\tilde{\mathbf{X}}_t$, it is crucial to be able to compare likelihoods when the number of objects is changing. Thus, we paid great care to propose a formulation that provides likelihoods of similar orders of magnitudes for different number of objects. For simplicity, we dropped the subscript t in this section. Our observations are defined as $\mathbf{Z} = (\mathbf{I}_v, \mathbf{D}_v)_{v=1..N_v}$, where $\mathbf{I}_v$ and $\mathbf{D}_v$ denotes the color and the background subtraction observations for each of the N_v camera views. More precisely, $\mathbf{D}_v$ is a background distance map obtained from the background subtraction of [12], with values between 0 and 1 where 0 means a perfect match with the background. Assuming the conditional independence of the camera views, we have:

$$p(\mathbf{Z}|\tilde{\mathbf{X}}) = \prod_{v=1}^{N_v} p(\mathbf{I}_v|\mathbf{D}_v, \tilde{\mathbf{X}})p(\mathbf{D}_v|\tilde{\mathbf{X}}). \tag{8}$$

These two terms are described below (where we dropped the subscript v for simplicity).

The Foreground Likelihood of one camera is modeled as:

$$p(\mathbf{D}|\tilde{\mathbf{X}}) = \prod_{\mathbf{x}\in S} \exp^{-\lambda_{fg}(1-\mathbf{D}(\mathbf{x}))} \prod_{\mathbf{x}\in \bar{S}} \exp^{-\bar{\lambda}_{fg}\mathbf{D}(\mathbf{x})} \propto \prod_{\mathbf{x}\in S} \exp^{c_1(\mathbf{D}(\mathbf{x})-c_2)} \tag{9}$$

where $\mathbf{x}$ denotes an image pixel, S denotes the object regions of the image, $\bar{S}$ denotes its complement, as illustrated in Fig. 3, and $c_1 = (\lambda_{fg} + \bar{\lambda}_{fg})$ and $c_2 = \lambda_{fg}/c_1$. In Eq. (9), we can clearly notice that the number of terms is independent of the number of objects, and that the placement (for track or birth) of objects will be encouraged in regions where $\mathbf{D}(\mathbf{x}) > c_2$.

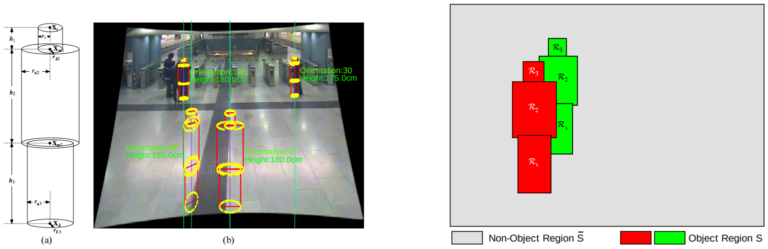


Figure 3: *Left: (a) the 3D human model consisting of three elliptic cylinders. (b) projection of the body model in the rectified image for different state values. Right: Object and non-object regions.*

The Color Likelihood in one camera is modeled as:

$$\begin{aligned} p(\mathbf{I}|\mathbf{D},\tilde{\mathbf{X}}) &= \prod_{i\in\mathcal{K}}\prod_{b=1}^{3} \exp^{-\lambda_{im}|\mathcal{R}_{i,b}|D_c(\mathbf{I},\mathbf{D},\mathcal{R}_{i,b})} \prod_{\mathbf{x}\in\bar{S}} \exp^{-\lambda_{im}D_{min}} \\ &\propto \prod_{i\in\mathcal{K}}\prod_{b=1}^{3} \exp^{-\lambda_{im}|\mathcal{R}_{i,b}|(D_c(\mathbf{I},\mathbf{D},\mathcal{R}_{i,b})-D_{min})} \end{aligned} \quad (10)$$

where $\mathcal{R}_{i,b}$ denotes, for an existing object i visible in the camera view, the image part of its body region b which are not covered by other objects (see Fig. 3), and $|\mathcal{R}_{i,b}|$ denotes the area of $\mathcal{R}_{i,b}$. The above expression provides a comparable likelihood for different number of objects, and will favor the placement of tracked objects at positions for which the body region color distance $D_c(\mathbf{I},\mathbf{D},\mathcal{R}_{i,b})$ is high, and favor the object existence if this distance is (on average) higher than the expected minimum distance D_{min}.

Object color representation and distance. From the visible part of the body region $\mathcal{R}_b$ of an object, we extract two color histograms: $\mathbf{h}_b$, which uses only foreground pixels (i.e. for which $\mathbf{D}(\mathbf{x}) > c_2$), and $\mathbf{H}_b$, which uses all pixels in $\mathcal{R}_b$. While the former should be more accurate by avoiding pooling pixels from the background, the latter one guarantees that we will have enough observations. To efficiently account for appearence variability due to pose, lighting, resolution and camera view changes, we propose to represent each object body region using a set of K automatically learned reference histograms, $\mathcal{H} = \{\bar{\mathbf{H}}_k\}_{k=1}^{K}$. The color distance is then defined as:

$$D_c(\mathbf{I},\mathbf{D},\mathcal{R}_b) = (1-\lambda_f)D_h^2\left(\mathbf{H}_b,\mathcal{H}\right) + \lambda_f D_h^2\left(\mathbf{h}_b,\mathcal{H}\right) \text{ with } D_h\left(\mathbf{H},\mathcal{H}\right) = \min_k D_{bh}\left(\mathbf{H},\bar{\mathbf{H}}_k\right)$$

where D_{bh} denotes the standard Bhattacharyya histogram distance. The updating of the reference histograms is conducted in a similar way to background modeling methods [12]: observed histograms (extracted from the mean object state) are matched against the reference histograms and used to update the best matched histogram, or create a new reference histogram if the best match is not close enough.

3.4 Reversible-Jump MCMC

Given the high and variable dimensionality of our state space, the inference of the filtering distribution $p(\tilde{\mathbf{X}}_t|\mathbf{Z}_{1:t})$ is conducted using a Reversible-Jump MCMC (RJ-MCMC) sampling scheme which has been shown to be the very efficient in such cases [4–6]. In RJ-MCMC, a Markov Chain is defined such that its stationary distribution is equal to the target distribution, Eq. (2) in our case. The Markov Chain is sampled using the Metropolis-Hastings (MH) algorithm. Starting from an arbitrary configuration, the algorithm proceeds by repetitively selecting a move type m from a set of moves Υ with prior probability p_m and sampling a new configuration $\tilde{\mathbf{X}}_t'$ from a proposal distribution $q_m(\tilde{\mathbf{X}}_t'|\tilde{\mathbf{X}}_t)$. The move can either change the dimensionality of the state (as in birth or

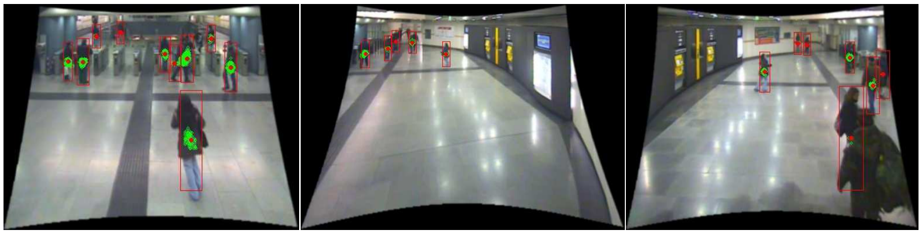


Figure 4: *Detection results at the same time instant in the 3 views.*

death) or keep it fixed. Then, either the proposed configuration is added with probability (known as *acceptance ratio*)

$$a = \frac{p(\tilde{\mathbf{X}}'_t|\mathbf{Z}_{1:t})}{p(\tilde{\mathbf{X}}_t|\mathbf{Z}_{1:t})} \times \frac{p_{m'}}{p_m} \times \frac{q_{m'}(\tilde{\mathbf{X}}_t;\tilde{\mathbf{X}}'_t)}{q_m(\tilde{\mathbf{X}}'_t;\tilde{\mathbf{X}}_t)} \tag{11}$$

to the Markov Chain, where m' is the reverse move of m, or the current configuration is added otherwise. More details on defining typical moves and corresponding acceptance ratios can be found in [4]. In the following, we describe the moves and proposals we used and highlight the key points.

Human Detection: Good and accurate automatic track initialization is crucial for multi-object tracking, in particular since it is the phase where the initial object model (color histograms) is extracted. In addition, being able to propose accurate positions to update current tracks is important. To this end, we have developed a human person detector [13] which builds on the approach of Tuzel *et al.* [14], and takes full advantage of the correlation existing between the shapes of humans in foreground detection maps and their appearance in the RGB images In multi-view calibrated environment, the detector was applied on each view separately, on windows i) which correspond to plausible people sizes; ii) for which the corresponding windows in the other camera views (obtained thanks to the calibration) all contained enough (20%) foreground pixels. Note that appart from this latter constraint, we did not try to merge the detection output in the different views. The main reason is that such fusion could reduce the number of detection (e.g. as the object might be too small, occluded or noisy in a given image). Also it appeared to be better to keep the best localizations in each of the camera views when initializing or updating track states in the MCMC tracking framework. Fig. 4 provides an example of obtained detections.

Move Proposals. We have defined six move types: *add*, *delete*, *stay*, *leave*, *switch*, and *update*. The proposal of each move type is defined as follows.

add/*delete*. the *add* move uses the output of the human detector, and proposes to randomly add one of the detected humans whose positions are far enough from the existing objects in the current configuration (where the distance is measured on the ground plane using the uncertainty Mahalanobis distance, cf Section 3.2 or Fig. 2). The *delete* move is the reverse move of the *add* move (reverse moves are required to potentially move the chain back to a previous hypothesis). In this move, objects which have been added with the *add* move are randomly selected for being removed.

stay/*leave*. the *add*/*delete* moves enable *new* objects to enter the scene, and is driven by a human detector. The *stay*/*leave* moves are the equivalent of *add*/*delete*, but allows to decide on the fate of objects that were already present at the previous time instant. The

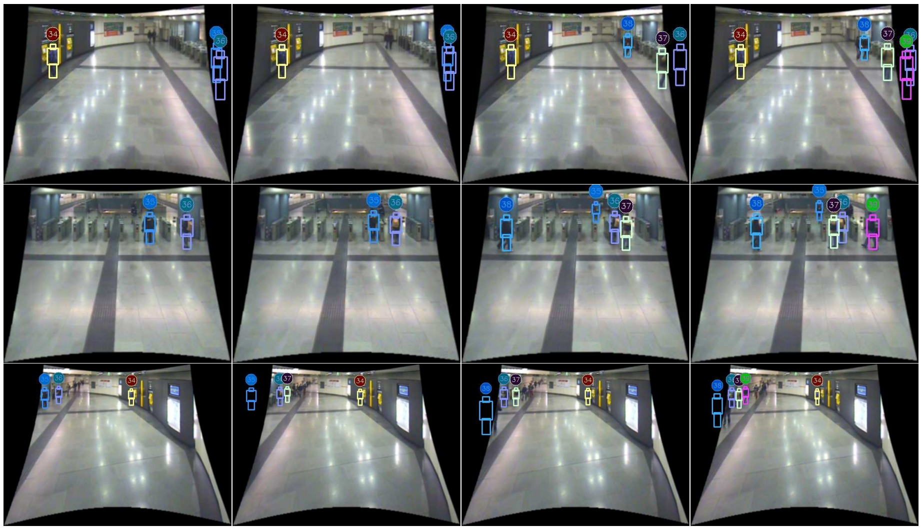


Figure 5: *Tracking results on the metro scene.*

leave move allows to remove one such object from the current configuration, while the *stay* allows to bring it back, by sampling from the state dynamics [4].

switch: This move allows to randomly exchange states between close-by objects, which in practice allows to check whether the exchange of color models better fits the data.

update: This is an important move which allows to find good estimates for the object states. It works by first randomly selecting a valid object i^* from the current joint configuration (i.e. for which $k_{i^*,t} = 1$), and then propose a new state for update. This new state is drawn in two ways (i.e the proposal is a mixture). In the first case, the object position, height and orientation are locally perturbed according to a Gaussian kernel [4]. Importantly, in order to propose interesting state values that may have a visual impact, the noise covariance in position is defined as $\Sigma(\mathbf{u}_{i^*}) = C(\mathbf{u}_{i^*})C^\top(\mathbf{u}_{i^*})$, where $C(\mathbf{u}_{i^*})$ is the noise matrix in Eq. (7) which is used to define the noise covariance in Fig. 2. The second way is to update the object location by sampling the new location around one of the positions provided by the human detector which are close enough from the selected object i^*. Here again, closeness is defined by exploiting $\Sigma(\mathbf{u}_{i^*})$, and the perturbation covariance around the selected detection is given by $\Sigma(\mathbf{u}_{i^*})$.

4 Experimental Results

Two datasets captured from two different scenes were used to evaluate our proposed multi-person tracking system. The first one consists of three 2h30 minutes video footage captured by three wide-baseline cameras in the Torino metro station scene as shown in Fig. 4. These sequences are very challenging, due to the camera view points (small average people size and large people size variations in a given view, occlusion, partial field of view overlap), crowded scenes in front of the gates, and the presence of many 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 second dataset comprises 10 minutes of video footage also captured by three wide-baseline cameras in an outdoor scene. In this scene, people often appear slanted in the left or/and right borders of an image (see Fig. 1). The camera view point

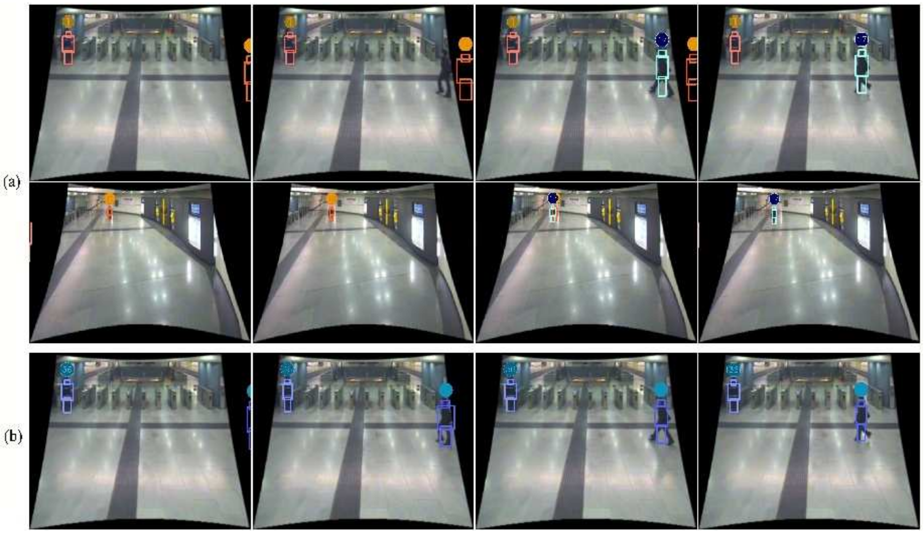


Figure 6: *Tracking results on the metro scene: (a) without integrating the ground plane noise model in dynamical model; (b) with integration.*

issues mentioned above for the metro scene also exist in this outdoor scene. The following experiments were obtained using a total of 1500 moves in the RJ-MCMC sampling with 500 in the burn-in phase.

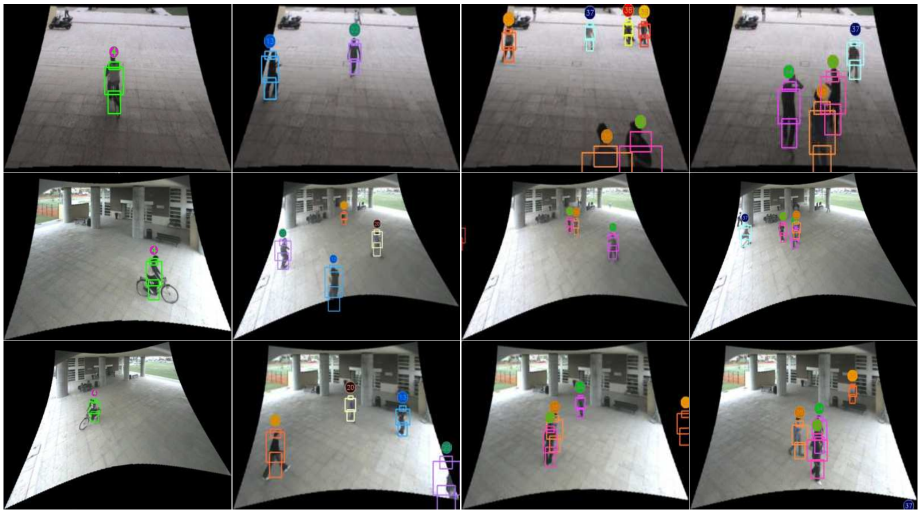
Fig. 5 shows some tracking results on the first dataset. In this example,our tracking system performed very well, successfully adding people using the human detector mediated birth move, and efficiently handling inter-person occlusion and partial visibility between camera views.

The benefit of using the 2D-to-3D ground plane noise in our algorithm, and especially in the dynamics, is illustrated in a simple example, Fig. 6. In the first two rows, this component was not used, i.e. $\mathbf{C}(\mathbf{u}) = 0$ in Eq. (7). As can be seen, the estimated state from the first view lags a little bit behind, resulting in a mismatch when the tracked person enters the second view. As a consequence, a new object is created. The first track stays for some time, and is then removed, resulting alltogether in a track break. On the other hand, when using the proposed term, the transition between cameras is successfully handled by the algorithm, as shown in Fig. 6(b).

On the second dataset, our approach performed very well, with almost no tracking errors in the 10-minute sequences. Results on four frames are shown in Fig. 7. Anecdotically, our human detector was able to successfully detect a person on a bicycle and our tracking system was able to track him/it robustly.

5 Conclusions

In this paper, we presented a novel multi-camera multi-person 3D tracking algorithm. The strength of the approach relies on several key factors: the joint multi-state Bayesian formulation, appropriate interaction models using state-prediction to model collision avoidance, the RJ-MCMC inference sampling scheme, and well balanced observation models. The use of a fast and powerful human detector proved to be essential for good track initialization and state update, as was the use of predefined 2D to 3D geometric uncertainty measures on the state dynamics. In addition, a novel simple rectification scheme was proposed to remove people slant from images and allow the use of efficient



Frame 197 Frame 973 Frame 1267 Frame 1288

Figure 7: *Example of multi-preson tracking on the outdoor sequence.*

human detector and feature extraction based on integral images. Future work is oriented towards the definition of more powerful learned object likelihood models, esp. to handle partial occlusion, on the use of longer term constraints on the dynamics.

References

1. Isard, M., MacCormick, J.: Bramble: A bayesian multi-blob tracker. In: Int. Conf. Comp. Vision (ICCV). (2001)
2. Tao, H., Sawhney, H., Kumar., R.: A sampling algorithm for detection and tracking multiple objects. In: ICCV Workshop on Vision Algorithms. (1999)
3. Tweed, D., Calway, A.: Tracking many objects using subordinate condensation. In: Europe Conf. Comp. Vision (ECCV). (2002)
4. Khan, Z.: Mcmc-based particle filtering for tracking a variable number of interacting targets. IEEE Trans. Pattern Anal. Machine Intell. **27** (2005) 1805–1918
5. Zhao, T., Nevatia, R.: Tracking multiple humans in complex situations. IEEE Trans. Pattern Anal. Machine Intell. **26** (2004) 1208–1221
6. Smith, K., Gatica-Perez, D., Odobez, J.: Using particles to track varying numbers of interacting people. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2005)
7. Pham, N.T., Huang, W., Ong, S.H.: Probability hypothesis density approach for multi-camera multi-object tracking. In: Asian Conf on Comp. Vision (ACCV). (2007)
8. Du, W., Piater, J.: Multi-camera people tracking by collaborative particle filters and principal axis-based integration. In: Asian Conf on Comp. Vision (ACCV). (2007)
9. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Trans. Pattern Anal. Machine Intell. **30** (2008) 267–282
10. Wang, G., Hu, Z., Wu, F., Tsui, H.T.: Single view metrology from scene constraints. Image & Vision Computing Journal **23** (2005) 831–840
11. Yao, J., Odobez, J.M.: Multi-camera 3d single person tracking with particle filter in a surveillance environment. In: 16th European Signal Processing Conference. (2008)
12. Yao, J., Odobez, J.M.: Multi-layer background subtraction based on color and texture. In: CVPR Visual Surveillance Workshop (VS-CVPR). (2007) 1–8
13. Yao, J., Odobez, J.M.: Fast human detection from videos using covariance features. In: ECCV 2008 Visual Surveillance Workshop. (2008)
14. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on riemannian manifolds. In: IEEE Conf. Comp. Vision & Pattern Recognition (CVPR). (2007)