An On-line Variational Bayesian Model for Multi-Person Tracking from Cluttered Scenes

Sileye Ba, Xavier Alameda-Pineda, Alessio Xompero, Radu Horaud

An On-line Variational Bayesian Model for Multi-Person Tracking from Cluttered Scenes

Sileye Ba[a,*], Xavier Alameda-Pineda[a,b], Alessio Xompero[a], Radu Horaud[a]

[a] *INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France*
[b] *University of Trento, Trento, Italy*

Abstract

Object tracking is an ubiquitous problem that appears in many applications such as remote sensing, audio processing, computer vision, human-machine interfaces, human-robot interaction, etc. Although thoroughly investigated in computer vision, tracking a time-varying number of persons remains a challenging open problem. In this paper, we propose an on-line variational Bayesian model for multi-person tracking from cluttered visual observations provided by person detectors. The paper has the following contributions. We propose a variational Bayesian framework for tracking an unknown and varying number of persons. Our model results in a variational expectation-maximization (VEM) algorithm with closed-form expressions both for the posterior distributions of the latent variables and for the estimation of the model parameters. The proposed model exploits observations from multiple detectors, and it is therefore multimodal by nature. Finally, we propose to embed both object-birth and object-visibility processes in an effort to robustly handle temporal appearances and disappearances. Evaluated on classical multiple person tracking datasets, our method shows competitive results with respect to state-of-the-art multiple-object tracking algorithms, such as the probability hypothesis density (PHD) filter, among others.

Keywords: Multi-person tracking, Bayesian tracking, variational expectation-maximization, causal inference, person detection.

1. Introduction

The problem of tracking a varying number of objects is ubiquitous in a number of fields such as remote sensing, computer vision, human-computer interaction, human-robot interaction, etc. While several off-line multi-object tracking methods are available, on-line multi-person tracking is still extremely challenging [1]. In this paper we propose an on-line tracking method within the tracking-by-detection (TbD) paradigm, which gained popularity in the computer vision community thanks to the development of efficient and robust object detectors [2]. Moreover, one advantage of TbD paradigm is the possibility of using linear mappings to link the kinematic (latent) states of the tracked objects to the observations issued from the detectors. This is possible because object detectors efficiently capture and filter out the non-linear effects, thus delivering detections that are linearly related to the kinematic latent states.

In addition to the difficulties associated to single-object tracking (occlusions, self-occlusions, visual appearance variability, unpredictable temporal behavior, etc.), tracking a varying and unknown number of objects makes the problem more challenging because of the following reasons: (i) the observations coming from detectors need to be associated to the objects that generated them, which includes the process of discarding detection errors, (ii) the number of objects is not known in advance and hence it must be estimated, mutual occlusions (not present in single-tracking scenarios) must be robustly handled, (iv) when many objects are present the dimension of the state-space is large, and hence the tracker has to handle a large number of hidden-state parameters, (v) the number of objects varies over time and one has to deal with hidden states of varying dimensionality, from zero when there is no visible object, to a large number of detected objects. Note that in this case and if a Bayesian setting is being considered, as is often the case, the exact recursive filtering solution is intractable.

In computer vision, previously proposed methodological frameworks for multi-target tracking can be divided into three groups. Firstly, the trans-dimensional Markov chain model [3], where the dimensionality of the hidden state-space is part of the state variable. This allows to track a variable number of objects by jointly estimating the number of objects and their kinematic states. In a computer vision scenario, [4, 5, 6] exploited this framework for tracking a varying number of objects. The main drawback is that the states are inferred by means of a reversible jump Markov chain Monte Carlo sampling, which is computationally expensive [7]. Secondly, a random finite set multi-target tracking formulation was proposed [8, 9, 10]. Initially used for radar applications [8], in this framework the targets are modeled as realizations of a random finite set which is composed of an unknown number of elements. Because an exact solution to this model is computationnally intensive, an approximation known as the probability hypothesis density (PHD) filter was proposed [11]. Further sampling-based approximations of random det based filters were subsequently proposed, e.g. [12, 13, 14]. These were exploited in [15] for tracking a time-varying number of active speakers using auditory cues

*Corresponding author
Email address: Sileye.Ba@inria.fr (Sileye Ba)

and in [16] for multi-target tracking using visual observations. Thirdly, conditional random fields (CRF) were also chosen to address multi-target tracking [17, 18, 19]. In this case, tracking is casted into an energy minimization problem. In another line of research, in radar tracking, other popular multi-targets tracking model are joint probabilistisc data assocation (JPDA), and multiple hypothesis filters [20].

In this paper we propose an on-line variational Bayesian framework for tracking an unknown and varying number of persons. Although variational model are very popular in machine learning, their use in computer vision for object tracking has been limited to tracking situation involving a fixed number of targets [21]. Variational Bayes methods approximate the joint a posteriori distribution of the latent variables by a separable distribution [22, 23]. In an on-line tracking scenario, where only causal (past) observations can be used, this translates into approximating the filtering distribution. This is in strong contrast with off-line trackers that use both past and future observations. The proposed tracking algorithm is therefore modeling the a posteriori distribution of the hidden states given all past observations. Importantly, the proposed framework leads to closed-form expressions for the posterior distributions of the hidden variables and for the model parameters, thus yielding an intrinsically efficient filtering procedure implemented via an variational EM (VEM) algorithm. In addition, a *clutter target* is defined so that spurious observations, namely detector failures, are associated to this target and do not contaminate the filtering process. Furthermore, our formalism allows to integrate in a principled way heterogeneous observations coming from various detectors, e.g, face, upper-body, silhouette, etc. Remarkably, objects that come in and out of the field of view, namely object appearance and disappearance, are handled by object birth and visibility processes. In details, we replace the classical death process by a visibility process which allows to put asleep tracks associated with persons that are no longer visible. The main advantage is that these tracks can be awaken as soon as new observations match their appearance with high confidence. Summarizing, the paper contributions are:

- Cast the problem of tracking a time-varying number of people into a variational Bayes formulation, which approximates the a posteriori filtering distribution by a separable distribution;
- A VEM algorithm with closed-form expressions, thus inherently efficient, for the update of the a posteriori distributions and the estimation of the model parameters from the observations coming from different detectors;
- An object-birth and an object-visibility process allowing to handle person appearance and disappearance due either to occlusions or people leaving the visual scene;
- A thorough evaluation of the proposed method compared with the state-of-the-art in two datasets, the cocktail party dataset and a dataset containing several sequences traditionally used in the computer vision community to evaluate multi-person trackers.

The remainder of this paper is organized as follows. Section 2 reviews previous work relevant to our work method. Section 3 details the proposed Bayesian model and a variational model solution is presented in Section 4. In Section 5, we depict the birth and visibility processes allowing to handle an unknown and varying number of persons. Section 6 describes results of experiments and benchmarks to assess the quality of the proposed method. Finally, Section 7 draws some conclusions.

2. Related Work

Generally speaking, object tracking is the temporal estimation of the object's kinematic state. In the context of image-based tracking, the object state is typically a parametrization of its localization in the (2D) image plane. In computer vision, object tracking has been thoroughly investigated [24]. Objects of interest could be people, faces, hands, vehicles, etc. According to the considered number of objects to be tracked, tracking can be classified into single-object tracking, fixed-number multi-object tracking, and varying-number multi-object tracking.

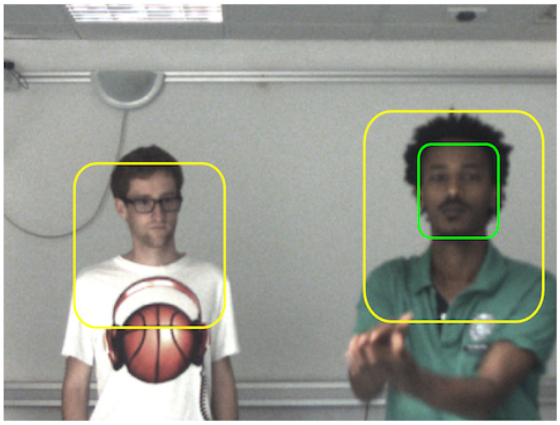
Methods for single object tracking consider only one object and usually involve an initialization step, a state update step, and a reinitialization step allowing to recover from failures. Practical initialization steps are based on generic object detectors allowing to scan the input image in order to find the object of interest [25, 26]. Object detectors can be used for the reinitialization step as well. However, using generic object detectors is problematic when other objects of the same kind than the tracked object are present in the visual scene. In order to resolve such ambiguities, different complementary appearance models have been proposed, such as object templates, color appearance models, edges (image gradients) and texture, (e.g. Gabor features and histogram of gradient orientations). Regarding the update step, the current state can be estimated from previous states and observations with either deterministic [27] or probabilistic [28] methods.

Even if it is still a challenging problem, tracking a single object is very limited in scope. Rapidly, the computer vision community drove its attention towards fixed-number multi-object tracking [29]. Additional difficulties are encountered when tracking multiple objects. Firstly, there is an increase of the tracking state dimensionality as the multi-object tracking state dimensionality is the single object state dimensionality multiplied by the number of tracked objects. Secondly, associations between observations and objects are required. Since the observation-to-object association problem is combinatorial [30, 20], it must be carefully addressed when the number of objects and of observations are large. Thirdly, because of the presence of multiple targets, tracking methods have to be robust also to mutual occlusions.

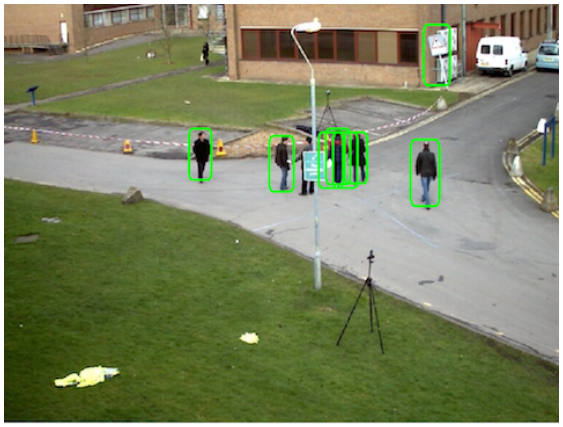
In most practical applications, the number of objects to be tracked, is not only unknown, but it also varies over time. Importantly, tracking a time-varying number or objects requires an efficient mechanism to add new objects entering the field of view, and to remove objects that moved away. In a probabilistic setting, these mechanisms are based on birth and death

processes. Efficient multi-object algorithms have to be developed within principled methodologies allowing to handle hidden states of varying dimensionality. In computer vision, the most popular methods are based on conditional random fields [31, 18, 19, 32], on random finite sets [10, 15, 16] or on the trans-dimensional Markov chain [3, 4, 5, 6]. [6] presents an interesting approach where occlusion state of a tracked person is explicitly modeled in the tracked state and used for observation likelihood computation. Less popular but successful methodologies include the Bayesian multiple blob tracker of [33], the boosted particle filter for multi-target tracking of [34] and the Rao-Blackwellized filter for multiple objects tracking [35], graph based representation for multi-object tracking [36, 37]. It has to be noticed in other communities, such as radar tracking, multi-object tracking has been deeply investigated. Many models have been proposed such as the probabilistic data association filter (PDAF), the joint PDAF, multiple hypothesis tracking [20]. However, the differences between multi-object tracking in radar and in computer vision are mainly two. On the one hand, most tracking method for radar consider point-wise objects, modeling a punctual latent state, whereas in computer vision objects are represented using bounding boxes in addition to the punctual coordinates. On the other hand, computer vision applications benefit from the use of visual appearance, which is mainly used for object identification [38].

Currently available multi-object tracking methods used in computer vision applicative scenarios suffer from different drawbacks. CRF-based approaches are naturally non-causal, that is, they use both past and future information. Therefore, even if they have shown high robustness to clutter, they are only suitable for off-line applications when smoothing (as opposite to filtering) techniques can be used. PHD filtering techniques report good computational efficiency, but they are inherently limited since they provide non-associated tracks. In other words, these techniques require an external method in order to associate observations and tracks to objects. Finally, even if trans-dimensional MCMC based tracking techniques are able to associate tracks to objects using only causal information, they are extremely complex from a computational point of view, and their performance is very sensitive to the sampling procedure used. In contrast, the variational Bayesian framework we propose associates tracks to previously seen objects and creates new tracks in an unified framework that filters past observations in an intrinsically efficient way, since all the steps of the algorithm are expressed in closed-form. Hence the proposed method robustly and efficiently tracks a varying and unknown number of persons from a combination of image detectors.



(a)



(b)

Figure 1: Examples of detections used as observations by the proposed person tracker: upper-body, face (a), and full-body (b) detections. Notice that one of the faces was not detected and that there is a false full-body detection in the background.

3. Variational Bayesian Multiple-Person Tracking

3.1. Notations and Definitions

We start by introducing our notations. Vectors and matrices are in bold $\mathbf{A}$, $\mathbf{a}$, scalars are in italic A, a. In general random variables are denoted with upper-case letters, e.g. $\mathbf{A}$ and A, and their realizations with lower-case letters, e.g. $\mathbf{a}$ and a.

Since the objective is to track multiple persons whose number may vary over time, we assume that there is a maximum number of people, denoted by N, that may enter the visual scene. This parameter is necessary in order to cast the problem at hand into a finite-dimensional state space, consequently N can be arbitrarily large. A track n at time t is associated to the *existence* binary variable e_{tn} taking the value $e_{tn} = 1$ if the person has already been seen and $e_{tn} = 0$ otherwise. The vectorization of the existence variables at time t is denoted by $\mathbf{e}_t = (e_{t1}, ..., e_{tN})$ and their sum, namely the effective number of tracked persons at t, is denoted by $N_t = \sum_{n=1}^{N} e_{tn}$. The existence variables are assumed to be observed in sections 3 and 4; Their inference, grounded in a track-birth stochastic process, is discussed in section 5.

The kinematic state of person n is a random vector $\mathbf{X}_{tn} = (\mathbf{L}_{tn}^\top, \mathbf{U}_{tn}^\top)^\top \in \mathbb{R}^6$, where $\mathbf{L}_{tn} \in \mathbb{R}^4$ is the person location, i.e., 2D image position, width and height, and $\mathbf{U}_{tn} \in \mathbb{R}^2$ is the person velocity in the image plane. The multi-person state random vector is denoted by $\mathbf{X}_t = (\mathbf{X}_{t1}^\top, \ldots, \mathbf{X}_{tN}^\top)^\top \in \mathbb{R}^{6N}$. Importantly, the kinematic state is described by a set of hidden variables which must be robustly estimated.

In order to ease the challenging task of tracking multiple per-

sons with a single static camera, we assume the existence of I detectors, each of them providing K_t^i localization observations at each time t, with $i \in [1 \dots I]$. Fig. 1 provides examples of face and upper-body detections (see Fig. 1(a)) and of full-body detections (see Fig. 1(b)). The k-th localization observation gathered by the i-th detector at time t is denoted by $\mathbf{y}_{tk}^i \in \mathbb{R}^4$, and represents the location (2D position, width, height) of a person in the image. The set of observations provided by detector i at time t is denoted by $\mathbf{y}_t^i = \{\mathbf{y}_{tk}^i\}_{k=1}^{K_t^i}$, and the observations provided by all the detectors at time t is denoted by $\mathbf{y}_t = \{\mathbf{y}_t^i\}_{i=1}^I$. Associated to each localization detection $\mathbf{y}_{tk}^i$, there is a photometric description of the person's appearance, denoted by $\mathbf{h}_{tk}^i$. This photometric observation is extracted from the bounding box of $\mathbf{y}_{tk}^i$. Altogether, the localization and photometric observations constitute the raw observations $\mathbf{o}_{tk}^i = (\mathbf{y}_{tk}^i, \mathbf{h}_{tk}^i)$ used by our tracker. Analogous definitions to $\mathbf{y}_t^i$ and $\mathbf{y}_t$ hold for $\mathbf{h}_t^i = \{\mathbf{h}_{tk}^i\}_{k=1}^{K_t}$, $\mathbf{h}_t = \{\mathbf{h}_t^i\}_{i=1}^I$, $\mathbf{o}_t^i = \{\mathbf{o}_{tk}^i\}_{k=1}^{K_t}$ and $\mathbf{o}_t = \{\mathbf{o}_t^i\}_{i=1}^I$. Importantly, when we write the probability of a set of random variables, we refer to the joint probabilities of all random variables in that set. For instance: $p(\mathbf{o}_t^i) = p(\mathbf{o}_{t1}^i, \dots, \mathbf{o}_{tK_t^i}^i)$.

We also define an observation-to-person assignment (hidden) variable Z_{tk}^i associated with each observation $\mathbf{o}_{tk}^i$. Formally, Z_{tk}^i is a categorical variable taking values in the set $\{1 \dots N\}$: $Z_{tk}^i = n$ means that $\mathbf{o}_{tk}^i$ is associated to person n. $\mathbf{Z}_t^i$ and $\mathbf{Z}_t$ are defined in an analogous way to $\mathbf{y}_t^i$ and $\mathbf{y}_t$. These assignment variables can be easily used to handle false detections. Indeed, it is common that a detection corresponds to some clutter instead of a person. We cope with these false detections by defining a *clutter* target. In practice, the index $n = 0$ is assigned to this clutter target, which is always visible, i.e. $e_{t0} = 1$ for all t. Hence, the set of possible values for Z_{tk}^i is extended to $\{0\} \cup \{1 \dots N\}$, and $Z_{tk}^i = 0$ means that observation $\mathbf{o}_{tk}^i$ has been generated by clutter and not by a person. The practical consequence of adding a clutter track is that the observations assigned to it play no role in the estimation of the parameters of the other tracks, thus leading to estimation rules inherently robust to outliers.

3.2. The Proposed Bayesian Multi-Person Tracking Model

The on-line multi-person tracking problem is cast into the estimation of the filtering distribution of the hidden variables given the causal observations $p(\mathbf{Z}_t, \mathbf{X}_t | \mathbf{o}_{1:t}, \mathbf{e}_{1:t})$, where $\mathbf{o}_{1:t} = \{\mathbf{o}_1, \dots, \mathbf{o}_t\}$. The filtering distribution can be rewritten as:

$$p(\mathbf{Z}_t, \mathbf{X}_t | \mathbf{o}_{1:t}, \mathbf{e}_{1:t}) = \frac{p(\mathbf{o}_t | \mathbf{Z}_t, \mathbf{X}_t, \mathbf{o}_{1:t-1}, \mathbf{e}_{1:t}) p(\mathbf{Z}_t, \mathbf{X}_t | \mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})}{p(\mathbf{o}_t | \mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})}. \tag{1}$$

Importantly, we assume that the observations at time t only depend on the hidden and visibility variables at time t. Therefore (1) writes:

$$p(\mathbf{Z}_t, \mathbf{X}_t | \mathbf{o}_{1:t}, \mathbf{e}_{1:t}) = \frac{p(\mathbf{o}_t | \mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t) p(\mathbf{Z}_t | \mathbf{e}_t) p(\mathbf{X}_t | \mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})}{p(\mathbf{o}_t | \mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})}. \tag{2}$$

The denominator of (2) only involves observed variables and therefore its evaluation is not necessary as long as one can normalize the expression arising from the numerator. Hence we focus on the three terms of the latter, namely the observation model $p(\mathbf{o}_t | \mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t)$, the observation-to-person assignment prior distribution $p(\mathbf{Z}_t | \mathbf{e}_t)$ and the dynamics of the latent state $p(\mathbf{X}_t | \mathbf{X}_{t-1}, \mathbf{e}_t)$, which appear when marginalizing the predictive distribution $p(\mathbf{X}_t | \mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})$ with respect to $\mathbf{X}_{t-1}$. Figure 2 shows a graphical schematic representation of the proposed probabilistic model.

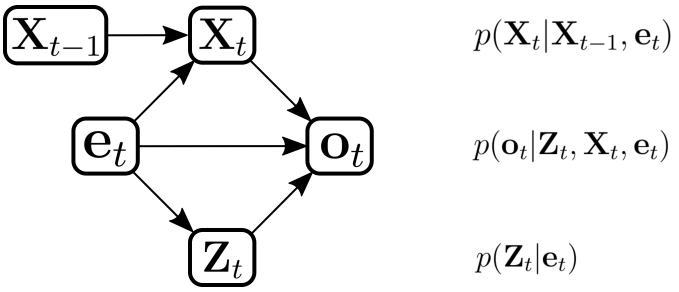


Figure 2: Graphical representation of the proposed multi-target tracking probabilistic model.

3.2.1. The Observation Model

The joint observations are assumed to be independent and identically distributed:

$$p(\mathbf{o}_t | \mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t) = \prod_{i=1}^{I} \prod_{k=1}^{K_t^i} p(\mathbf{o}_{tk}^i | Z_{tk}^i, \mathbf{X}_t, \mathbf{e}_t). \tag{3}$$

In addition, we make the reasonable assumption that, while localization observations depend both on the assignment variable and kinematic state, the appearance observations only depend on the assignment variable, that is the person identity, but not on his/her kinematic state. We also assume the localization and appearance observations to be independent given the hidden variables. Consequently, the observation likelihood of a single joint observation can be factorized as:

$$\begin{aligned} p(\mathbf{o}_{tk}^i | Z_{tk}^i, \mathbf{X}_t, \mathbf{e}_t) &= p(\mathbf{y}_{tk}^i, \mathbf{h}_{tk}^i | Z_{tk}^i, \mathbf{X}_t, \mathbf{e}_t) \\ &= p(\mathbf{y}_{tk}^i | Z_{tk}^i, \mathbf{X}_t, \mathbf{e}_t) p(\mathbf{h}_{tk}^i | Z_{tk}^i, \mathbf{e}_t). \end{aligned} \tag{4}$$

The localization observation model is defined depending on whether the observation is generated by clutter or by a person:

- If the observation is generated from clutter, namely $Z_{tk}^i = 0$, the variable $\mathbf{y}_{tk}^i$ follows an uniform distribution with probability density function $u(\mathbf{y}_{tk}^i)$;
- If the observation is generated by person n, namely $Z_{tk}^i = n$, the variable $\mathbf{y}_{tk}^i$ follows a Gaussian distribution with mean $\mathbf{P}^i \mathbf{X}_{tn}$ and covariance $\mathbf{\Sigma}^i$: $\mathbf{y}_{tk}^i \sim g(\mathbf{y}_{tk}^i; \mathbf{P}^i \mathbf{X}_{tn}, \mathbf{\Sigma}^i)$

The linear operator $\mathbf{P}^i$ maps the kinematic state vectors onto the i-th space of observations. For example, when $\mathbf{X}_{tn}$ represents the upper-body kinematic state (upper-body localization and velocity) and $\mathbf{y}_{tk}^i$ represents the upper-body localization observation, $\mathbf{P}^i$ is a projection which, when applied to a state vector, only retains the localization components of the state vector. When $\mathbf{y}_{tk}^i$ is a face localization observation, the operator

$\mathbf{P}^i$ is a composition of the previous projection, and an affine transformation mapping an upper-body bounding-box onto its corresponding face bounding-box. Finally, the full observation model is compactly defined by

$$p(\mathbf{y}^i_{tk}|Z^i_{tk}=n,\mathbf{X}_t,\mathbf{e}_t)=u(\mathbf{y}^i_{tk})^{1-e_{tn}}\left(u(\mathbf{y}^i_{tk})^{\delta_{0n}}g(\mathbf{y}^i_{tk};\ \mathbf{P}^i\mathbf{X}_{tn},\mathbf{\Sigma}^i)^{1-\delta_{0n}}\right)^{e_{tn}}, \tag{5}$$

where δ_{ij} stands for the Kronecker function.

The appearance observation model is also defined depending on whether the observations is clutter or not. When the observation is generated by clutter, the appearance observation follows a uniform distribution with density function $u(\mathbf{h}^i_{tk})$. When the observation is generated by person n, the appearance observation follows a Bhattacharya distribution with density defined as

$$b(\mathbf{h}^i_{tk};\mathbf{h}_n)=\frac{1}{W_\lambda}\exp(-\lambda d_B(\mathbf{h}^i_{tk},\mathbf{h}_n)),$$

where λ is a positive skewness parameter, $d_B(.,.)$ is the Battacharya distance between histograms, $\mathbf{h}_n$ is the n-th person's reference appearance model[1]. This gives the following compact appearance observation model:

$$p(\mathbf{h}^i_{tk}|Z^i_{tk}=n,\mathbf{X}_t,\mathbf{e}_t)=u(\mathbf{h}^i_{tk})^{1-e_{tn}}(u(\mathbf{h}^i_{tk})^{\delta_{0n}}b(\mathbf{h}^i_{tk};\mathbf{h}_n)^{1-\delta_{0n}})^{e_{tn}}. \tag{6}$$

3.2.2. The Observation-to-Person Prior Distribution

The joint distribution of the assignment variables factorizes as:

$$p(\mathbf{Z}_t|\mathbf{e}_t)=\prod_{i=1}^{I}\prod_{k=1}^{K^i_t}p(Z_{tk}|\mathbf{e}_t). \tag{7}$$

When observations are not yet available, given existence variables $\mathbf{e}_t$, the assignment variables Z^i_{tk} are assumed to follow multinomial distributions defined as:

$$p(Z^i_{tk}=n|\mathbf{e}_t)=e_{tn}a^i_{tn}\quad\text{with}\quad\sum_{n=0}^{N}e_{tn}a^i_{tn}=1. \tag{8}$$

Because e_{tn} takes the value 1 only for actual persons, the probability to assign an observation to a non-existing person is null.

3.2.3. The Predictive Distribution

The kinematic state predictive distribution represents the probability distribution of the kinematic state at time t given the observations up to time $t-1$ and the existence variables $p(\mathbf{X}_t|\mathbf{o}_{1:t-1},\mathbf{e}_{1:t})$. The predictive distribution is mainly driven by the dynamics of people's kinematic states, which are modeled consdering two hypothesis. Firstly the kinematic state dynamics follow a first-order Markov chain, meaning that the state $\mathbf{X}_t$ only depends on state $\mathbf{X}_{t-1}$. Secondly, the person locations do not influence each other's dynamics, meaning that there is one first-order Markov chain for each person. Formally, this can be written as:

$$p(\mathbf{X}_t|\mathbf{X}_{1:t-1},\mathbf{e}_{1:t})=p(\mathbf{X}_t|\mathbf{X}_{t-1},\mathbf{e}_t)=\prod_{n=1}^{N}p(\mathbf{X}_{tn}|\mathbf{X}_{t-1n},e_{tn}). \tag{9}$$

The immediate consequence is that the posterior distribution computes:

$$p(\mathbf{X}_t|\mathbf{o}_{1:t-1},\mathbf{e}_{1:t})=\int\left(\prod_{n=1}^{N}p(\mathbf{X}_{tn}|\mathbf{x}_{t-1n},e_{tn})\right)p(\mathbf{x}_{t-1}|\mathbf{o}_{1:t-1},\mathbf{e}_{1:t-1})d\mathbf{x}_{t-1}. \tag{10}$$

For the model to be complete, $p(\mathbf{X}_{tn}|\mathbf{X}_{t-1,n},e_{tn})$ needs to be defined. The temporal evolution of the kinematic state $\mathbf{X}_{tn}$ is defined as:

$$p(\mathbf{X}_{tn}=\mathbf{x}_{tn}|\mathbf{X}_{t-1,n}=\mathbf{x}_{t-1,n},e_{tn})=u(\mathbf{x}_{tn})^{1-e_{tn}}g(\mathbf{x}_{tn};\ \mathbf{D}\mathbf{x}_{t-1,n},\mathbf{\Lambda}_n)^{e_{tn}}, \tag{11}$$

where $u(\mathbf{x}_{tn})$ is a uniform distribution over the motion state space, g is a Gaussian probability density function, $\mathbf{D}$ represents the dynamics transition operator, and $\mathbf{\Lambda}_n$ is a covariance matrix accounting for uncertainties on the state dynamics. The transition operator is defined as:

$$\mathbf{D}=\begin{pmatrix}\mathbf{I}_{4\times4} & \mathbf{I}_{2\times2}\\ \mathbf{0}_{2\times4} & \mathbf{I}_{2\times2}\end{pmatrix}=\begin{pmatrix}1&0&0&0&1&0\\0&1&0&0&0&1\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix}.$$

In other words, the dynamics of an existing person n is *either* follows a Gaussian with mean vector $\mathbf{D}\mathbf{X}_{t-1,n}$ and covariance matrix $\mathbf{\Lambda}_n$, *or* a uniform distribution if person n does not exist. The complete set of parameters of the proposed model is denoted with $\mathbf{\Theta}=(\{\mathbf{\Sigma}^i\}_{i=1}^I,\ \{\mathbf{\Lambda}_n\}_{n=1}^N,\mathbf{A}_{1:t})$, with $\mathbf{A}_t=\{a^i_{tn}\}_{n=0,i=1}^{N,I}$.

4. Variational Bayesian Inference

Because of the combinatorial nature of the observation-to-person assignment problem, a direct optimization of the filtering distribution (2) with respect to the hidden variables is intractable. We propose to overcome this problem via a variational Bayesian inference method. The principle of this family of methods is to approximate the intractable filtering distribution $p(\mathbf{Z}_t,\mathbf{X}_t|\mathbf{o}_{1:t},\mathbf{e}_{1:t})$ by a separable distribution, e.g. $q(\mathbf{Z}_t)\prod_{n=0}^{N}q(\mathbf{X}_{tn})$. According to the variational Bayesian formulation [22, 23], given the observations and the parameters at the previous iteration $\mathbf{\Theta}^\circ$, the optimal approximation has the following general expression:

$$\log q(\mathbf{Z}_t)=\mathbf{E}_{q(\mathbf{X}_t)}\left\{\log p(\mathbf{Z}_t,\mathbf{X}_t|\mathbf{o}_{1:t},\mathbf{e}_{1:t},\mathbf{\Theta}^\circ)\right\}, \tag{12}$$

$$\log q(\mathbf{X}_{tn})=\mathbf{E}_{q(\mathbf{Z}_t)\prod_{m\neq n}q(\mathbf{X}_{tm})}\left\{\log p(\mathbf{Z}_t,\mathbf{X}_t|\mathbf{o}_{1:t},\mathbf{e}_{1:t},\mathbf{\Theta}^\circ)\right\}. \tag{13}$$

In our particular case, when these two equations are put together with the probabilistic model defined in (3), (7) and (9), the expression of $q(\mathbf{Z}_t)$ factorizes further into:

$$\log q(Z^i_{tk})=\mathbf{E}_{q(\mathbf{X}_t)}\left\{\log p(Z^i_{tk},\mathbf{X}_t|\mathbf{o}_{1:t},\mathbf{e}_{1:t},\mathbf{\Theta}^\circ)\right\}, \tag{14}$$

[1] It should be noted that the normalization constant $W_\lambda=\int_{\sum_k\mathbf{h}_k=1}\exp(-\lambda d_B(\mathbf{h},\mathbf{h}_n))\mathrm{d}\mathbf{h}$ can be exactly computed only for histograms with dimension lower than 3. In practice W_λ is approximated using Monte Carlo integration.

Note that this equation leads to a finer factorization that the one we imposed. This behavior is typical of variational Bayes methods in which a very mild separability assumption can lead to a much finer factorization when combined with priors over hidden states and latent variables, i.e. (3), (7) and (9). The final factorization writes:

$$p(\mathbf{Z}_t, \mathbf{X}_t|\mathbf{o}_{1:t}, \mathbf{e}_{1:t}) \approx \prod_{i=1}^{I}\prod_{k=0}^{K_t^i} q(Z_{tk}^i) \prod_{n=0}^{N} q(\mathbf{X}_{tn}). \quad (15)$$

Once the posterior distribution over the hidden variables is computed (see below), the optimal parameters are estimated using $\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} J(\boldsymbol{\Theta}, \boldsymbol{\Theta}^\circ)$ with J defined as:

$$J(\boldsymbol{\Theta}, \boldsymbol{\Theta}^\circ) = \mathbf{E}_{q(\mathbf{Z}_t, \mathbf{X}_t)} \left\{\log p(\mathbf{Z}_t, \mathbf{X}_t, \mathbf{o}_{1:t}|\mathbf{e}_{1:t}, \boldsymbol{\Theta}, \boldsymbol{\Theta}^\circ)\right\}. \quad (16)$$

To summarize, the proposed solution for multi-person tracking is an on-line variational EM algorithm. Indeed, the factorization (15) leads to a variational EM in which the E-step consists of computing (14) and (13) and the M-step consists of maximizing the expected complete-data log-likelihood (16) with respect to the parameters. However, as is detailed below, for stability reasons the covariance matrices are not estimated with the variational inference framework, but set to a fixed value. The expectation and maximization steps of the algorithm are now detailed.

4.1. E-Z-Step

The estimation of $q(Z_{tk}^i)$ is carried out by developing the expectation (14). More derivation details can be found in Appendix A.2, which yields the following formula:

$$q(Z_{tk}^i = n) = \alpha_{tkn}^i, \quad (17)$$

where

$$\alpha_{tkn}^i = \frac{e_{tn}\epsilon_{tkn}^i a_{tn}^i}{\sum_{m=0}^{N} e_{tm}\epsilon_{tkm}^i a_{tn}^i}, \quad (18)$$

and ϵ_{tkn}^i is defined as:

$$\epsilon_{tkn}^i = \begin{cases} u(\mathbf{y}_{tk}^i)u(\mathbf{h}_{tk}^i) & n = 0, \\ g(\mathbf{y}_{tk}^i, \mathbf{P}^i\boldsymbol{\mu}_{tn}, \boldsymbol{\Sigma}^i)e^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{P}^{i\top}(\boldsymbol{\Sigma}^i)^{-1}\mathbf{P}^i\boldsymbol{\Gamma}_{tn}\right)} b(\mathbf{h}_{tk}^i; \mathbf{h}_n) & n \neq 0, \end{cases} \quad (19)$$

where $\mathrm{Tr}(\cdot)$ is the trace operator and $\boldsymbol{\mu}_{tn}$ and $\boldsymbol{\Gamma}_{tn}$ are defined by (21) and (22) below. Intuitively, this approximation shows that the assignment of an observation to a person is based on spatial proximity between the observation localization and the person localization, and the similarity between the observation's appearance and the person's reference appearance.

4.2. E-X-Step

The estimation of $q(\mathbf{X}_{tn})$ is derived from (13). Similarly to the previous posterior distribution, the mathematical derivations are provided in Appendix A.3, and boil down to the following formula:

$$q(\mathbf{X}_{tn}) = u(\mathbf{X}_{tn})^{1-e_{tn}} g(\mathbf{X}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn})^{e_{tn}}, \quad (20)$$

where the mean vector $\boldsymbol{\mu}_{tn}$ and the covariance matrix $\boldsymbol{\Gamma}_{tn}$ are given by

$$\boldsymbol{\Gamma}_{tn} = \Big(\sum_{i=1}^{I}\sum_{k=0}^{K_t^i} \alpha_{tkn}^i \left(\mathbf{P}^{i\top}\left(\boldsymbol{\Sigma}^i\right)^{-1}\mathbf{P}^i\right) + (\mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_n)^{-1}\Big)^{-1} \quad (21)$$

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn}\Big(\sum_{i=1}^{I}\sum_{k=0}^{K_t^i} \alpha_{tkn}^i \mathbf{P}^{i\top}\left(\boldsymbol{\Sigma}^i\right)^{-1}\mathbf{y}_{tk}^i + (\mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_n)^{-1}\mathbf{D}\boldsymbol{\mu}_{t-1,n}\Big). \quad (22)$$

We note that the variational approximation of the kinematic-state distribution reminds the Kalman filter solution of a linear dynamical system with mainly one difference: in our solution (21) and (22), the mean vectors and covariance matrices are computed with the observations weighted by α_{tkn}^i (see (21) and (22)).

4.3. M-step

Once the posterior distribution of the hidden variables is estimated, the optimal parameter values can be estimated via maximization of J defined in (16). The M-step allows to estimate the model parameter.

Regarding the parameters of the a priori observation-to-object assignment $\mathbf{A}_t$ we compute:

$$J(a_{tn}^i) = \sum_{k=1}^{K_t^i} e_{tn}\alpha_{tkn}^i \log(e_{tn}a_{tn}^i) \quad \text{s.t.} \quad \sum_{n=0}^{N} e_{tn}a_{tn}^i = 1, \quad (23)$$

and trivially obtain:

$$a_{tn}^i = \frac{e_{tn}\sum_{k=1}^{K_t^i}\alpha_{tkn}^i}{\sum_{m=0}^{N} e_{tm}\sum_{k=1}^{K_t^i}\alpha_{tkm}^i}. \quad (24)$$

The M-Step for observation covariances corresponds to the estimation of $\boldsymbol{\Sigma}^i$. This is done by maximizing

$$J(\boldsymbol{\Sigma}^i) = \sum_{k=1}^{K_t^i}\sum_{n=1}^{N} e_{tn}\alpha_{tkn}^i \log(\mathbf{y}_{tk}^i, \mathbf{P}^i\mathbf{X}_{tn}, \boldsymbol{\Sigma}^i)$$

with respect to $\boldsymbol{\Sigma}^i$. Differentiating $J(\boldsymbol{\Sigma}^i)$ with respect to $\boldsymbol{\Sigma}^i$ and equating to zero gives:

$$\boldsymbol{\Sigma}^i = \frac{1}{K_t^i N}\sum_{k=1}^{K_t^i}\sum_{n=1}^{N} e_{tn}\alpha_{tkn}^i \left(\mathbf{P}^i\boldsymbol{\Gamma}_{tn}\mathbf{P}^{i\top} + (\mathbf{y}_{tk}^i - \mathbf{P}^i\boldsymbol{\mu}_{tn})(\mathbf{y}_{tk}^i - \mathbf{P}^i\boldsymbol{\mu}_{tn})^\top\right) \quad (25)$$

The M-Step for kinematic state dynamics covariances corresponds to the estimation of $\boldsymbol{\Lambda}_n$ for a fixed n. This done by maximizing cost function

$$J(\boldsymbol{\Lambda}_n) = \mathbf{E}_{q(\mathbf{X}_{tn}|e_{tn})}[\log g(\mathbf{X}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1n}, \mathbf{D}\boldsymbol{\Gamma}_{tn}\mathbf{D}^\top + \boldsymbol{\Lambda}_n)^{e_{tn}})].$$

Equating differential of the cost $J(\boldsymbol{\Lambda}_n)$ to zeros gives:

$$\boldsymbol{\Lambda}_n = \mathbf{D}\boldsymbol{\Gamma}_{t-1n}\mathbf{D}^\top + \boldsymbol{\Gamma}_{tn} + (\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1,n})(\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1,n})^\top \quad (26)$$

It is worth noticing that, in the current filtering formalism, the formulas for $\mathbf{\Sigma}^i$ and $\mathbf{\Lambda}_n$ are instantaneous, i.e., they are estimated only from the observations at time t. The information at time t is usually insufficient to obtain stable values for these matrices. Even if estimating $\mathbf{\Sigma}^i$ and $\mathbf{\Lambda}_n$ is suitable in a parameter learning scenario where the tracks are provided, we noticed that in practical tracking scenarios, where the tracks are unknown, this does not yield stable results. Suitable priors on the temporal dynamics of the covariance parameters are required. Therefore, in this paper we assume that the observation and dynamical model covariance matrices are fixed.

5. Person-Birth and Person-Visibility Processes

Tracking a time-varying number of targets requires procedures to create tracks when new targets enter the scene and to delete tracks when corresponding targets leave the visual scene. In this paper, we propose a statistical-test based birth process that creates new tracks and a hidden Markov model (HMM) based visibility process that handles disappearing targets. Until here, we assumed that the existence variables e_{tn} were given.In this section we present the inference modelfor the existence variable based on the stochastic birth-process.

5.1. Birth Process

The principle of the person birth process is to search for consistent trajectories in the history of observations associated to clutter. Intuitively, two hypotheses "*the considered observation sequence is generated by a person not being tracked*" and "*the considered observation sequence is generated by clutter*" are confronted.

The model of "*the considered observation sequence is generated by a person not being tracked*" hypothesis is based on the observations and dynamic models defined in (5) and (11). If there is a not-yet-tracked person n generating the considered observation sequence $\{\mathbf{y}_{t-L,k_L}, \ldots, \mathbf{y}_{t,k_0}\}$,[2] then the observation likelihood is $p(\mathbf{y}_{t-l,k_l}|\mathbf{x}_{t-l,n}) = g(\mathbf{y}_{t-l,k_l}; \mathbf{P}\mathbf{x}_{t-l,n}, \mathbf{\Sigma})$ and the person trajectory is governed by the dynamical model $p(\mathbf{x}_{t,n}|\mathbf{x}_{t-1,n}) = g(\mathbf{x}_{t,n}; \mathbf{D}\mathbf{x}_{t-1,n}, \mathbf{\Lambda}_n)$. Since there is no prior knowledge about the starting point of the track, we assume a "flat" Gaussian distribution over $\mathbf{x}_{t-L,n}$, namely $p_b(\mathbf{x}_{t-L,n}) = g(\mathbf{x}_{t-L,n}; \mathbf{m}_b, \mathbf{\Gamma}_b)$, which is approximatively equivalent to a uniform distribution over the image. Consequently, the joint observation distribution writes:

$$\begin{aligned}\tau_0 =& p(\mathbf{y}_{t,k_0}, \ldots, \mathbf{y}_{t-L,k_L}) \\ =& \int p(\mathbf{y}_{t,k_0}, \ldots, \mathbf{y}_{t-L,k_L}, \mathbf{x}_{t:t-L,n}) d\mathbf{x}_{t:t-L,n} \\ =& \int \prod_{l=0}^{L} p(\mathbf{y}_{t,k_l}|\mathbf{x}_{t-l,n}) \\ & \times \prod_{l=0}^{L-1} p(\mathbf{x}_{t-l,n}|\mathbf{x}_{t-l-1,n}) \times p_b(\mathbf{x}_{t-2,n}) d\mathbf{x}_{t:t-L,n}, \end{aligned} \tag{27}$$

which can be seen as the marginal of a multivariate Gaussian distribution. Therefore, the joint observation distribution $p(\mathbf{y}_{t,k_0}, \mathbf{y}_{t-1,k_1}, \ldots, \mathbf{y}_{t-2,k_L})$ is also Gaussian and can be explicitly computed.

The model of "*the considered observation sequence is generated by clutter*" hypothesis is based on the observation model given in (5). When the considered observation sequence $\{\mathbf{y}_{t,k_0}, \ldots, \mathbf{y}_{t-L,k_L}\}$ is generated by clutter, observations are independent and identically uniformly distributed. In this case, the joint observation likelihood is

$$\tau_1 = p(\mathbf{y}_{t,k_0}, \ldots, \mathbf{y}_{t-L,k_L}) = \prod_{l=0}^{L} u(\mathbf{y}_{t-l,k_l}). \tag{28}$$

Finally, our birth process is as follows: for all $\mathbf{y}_{t,k_0}$ such that $\tau_0 > \tau_1$, a new person is added by setting $e_{tn} = 1$, $q(\mathbf{x}_{t,n}; \boldsymbol{\mu}_{t,n}, \mathbf{\Gamma}_{t,n})$ with $\boldsymbol{\mu}_{t,n} = [\mathbf{y}_{t,k_0}^\top, \mathbf{0}_2^\top]^\top$, and $\mathbf{\Gamma}_{tn}$ is set to the value of a birth covariance matrix (see (20)). Also, the reference appearance model for the new person is defined as $\mathbf{h}_{t,n} = \mathbf{h}_{t,k_0}$.

5.2. Person-Visibility Process

A tracked person is said to be visible at time t whenever there are observations associated to that person, otherwise the person is considered not visible. Instead of deleting tracks, as classical for death processes, our model labels tracks without associated observations as *sleeping*. In this way, we keep the possibility to awake such sleeping tracks when their reference appearance model highly matches an observed appearance.

We denote the n-th person visibility (binary) variable by V_{tn}, meaning that the person is visible at time t if $V_{tn} = 1$ and 0 otherwise. We assume the existence of a transition model for the hidden visibility variable V_{tn}. More precisely, the visibility state temporal evolution is governed by the transition matrix, $p(V_{tn} = j|V_{t-1,n} = i) = \pi_v^{\delta_{ij}}(1 - \pi_v)^{1-\delta_{ij}}$, where π_v is the probability to remain in the same state. To enforce temporal smoothness, the probability to remain in the same state is taken higher than the probability to switch to another state.

The goal now is to estimate the visibility of all the persons. For this purpose we define the visibility observations as $\nu_{tn} = e_{tn} \sum_{i=1}^{I} a_{tn}^i$, being 0 when no observation is associated to person n. In practice, we need to filter the visibility state variables V_{tn} using the visibility observations ν_{tn}. In other words, we need to estimate the filtering distribution $p(V_{tn}|\nu_{1:tn}, e_{1:tn})$ which can be written as:

$$\begin{aligned} p(V_{tn} = \nu_{tn}|\nu_{1:t}, e_{1:tn}) = \\ \frac{p(\nu_{tn}|\nu_{tn}, e_{tn}) \sum_{\nu_{t-1,n}} p(\nu_{tn}|\nu_{t-1,n}) p(\nu_{t-1,n}|\nu_{1:t-1,n}, e_{1:t-1})}{p(\nu_{tn}|\nu_{1:t-1,n}, e_{1:t})}, \end{aligned} \tag{29}$$

where the denominator corresponds to integrating the numerator over ν_{tn}. In order to fully specify the model, we define the visibility observation likelihood as:

$$p(\nu_{tn}|\nu_{tn}, e_{tn}) = (\exp(-\lambda\nu_{tn}))^{\nu_{tn}}(1 - \exp(-\lambda\nu_{tn}))^{1-\nu_{tn}} \tag{30}$$

Intuitively, when ν_{tn} is high, the likelihood is large if $\nu_{tn} = 1$ (person is visible). The opposite behavior is found when ν_{tn} is

[2] In practice we considered $L = 2$, however, derivations are valid for arbitrary values of L.

small. Importantly, at each frame, because the visibility state is a binary variable, its filtering distribution can be straightforwardly computed.

6. Experiments

6.1. Evaluation Protocol

We experimentally assess the performance of the proposed model using two datasets. The cocktail party dataset (CPD) is composed of two videos, CPD-2 and CPD-3, recorded with a close-view camera (see Figure 3(a) and 3(b)). Only people's upper body is visible, and mutual occlusions happen often. CPD-3 records 3 persons during 853 frames and CPD-2 records 2 persons during 495 frames.

The second dataset is constituted of four sequences classically used in computer vision to evaluate multi-person tracking methods [18, 19]. Two sequences were selected from the MOT Challenge Dataset [39]:[3] TUD-Stadmitte (9 persons, 179 frames) and PETS09-S2L1 (18 persons, 795 frames). The third sequence is the TownCentre sequence (231 persons, 4500 frames) recorded by the Oxford Active Vision Lab. The last one is ParkingLot (14 persons, 749) recorded by the Center for Research in Computer Vision of University of Central Florida. TUD-Stadmitte records closely viewed full body pedestrians. PETS09-S2L1 and ParkingLot features a dozen of far-viewed full body pedestrians. TownCentre captures a very large number of far viewed pedestrians. This evaluation dataset is diverse and large (more than 6000 frames) enough to give a reliable assessment of the multi-person tracking performance measures. Figure 3 shows typical views of all the sequences.

Because multi-person tracking intrinsically implies track creation, deletion, target identity maintenance, and localization, evaluating multi-person tracking models is a non-trivial task. Many metrics have been proposed, see [40, 41, 42, 43]. In this paper, for the sake of completeness we use several of them split into two groups.

The first set of metrics follow the widely used CLEAR multi-person tracking evaluation metrics [42] which are commonly used to evaluate multi-target tracking where targets' identities are jointly estimated together with their kinematic states. On the one side the *multi-object tracking accuracy* (**MOTA**) combines false positives (FP), missed targets (FN), and identity switches (ID). On the other side, the *multi-object tracking precision* (**MOTP**) measures the alignment of the tracker output bounding box with the ground truth. We also provide tracking precision (**Pr**) and recall (**Rc**).

The second group of metrics is specifically designed for multi-target tracking models that do not estimate the targets' identities, such as the PHD filter. These metrics compute set distances between the ground truth set of objects present in the scene and the set of objects estimated by the tracker [40]. The metrics are the **Hausdorff** metric, the optimal mass transfer (**OMAT**) metric, and the optimal sub-pattern assignment (**OSPA**) metric. We will use these metrics to compare the tracking results achieved by our variational tracker to the results achieved by the PHD filter which does not infer identities [44].

The computational cost of the proposed model is mainly due the the observation extraction, namely the person detection. This process is known in computer vision to be computationally intensive. However, there are pedestrian detectors that achieve real time performances [45]. The VEM part of the tracking model, which involves only inversion of 6 by 6 matrices, is computationally efficient and can be made real time. It converges in less than 10 steps.

6.2. Validation on the Cocktail Party Dataset

In the cocktail party dataset our model exploits upper body detections obtained using [25] and face detections obtained using [26]. Therefore, we have two types of observations, upper body U and face F. The hidden state corresponds to the position and velocity of the upper body. The observation operator $\mathbf{P}^{\mathrm{U}}$ (see section 3.2.1) for the upper body observations simply removes the velocity components of the hidden state. The observation operator $\mathbf{P}^{\mathrm{F}}$ for the face observations combines a projection removing the velocity components and an affine mapping (scaling and translation) transforming face localization bounding boxes into the the upper body localization bounding boxes. The appearance observations are concatenations of joint hue-saturation color histograms of the torso split into three different regions, plus the head region as shown in Fig.4(a).

Tables 1 and 2 show the performance of the model over the two sequences of the cocktail party dataset. While in Table 1 we evaluate the performance of our model under the first set of metrics, in Table 2, we compare the performance of our model to the one of the GMM PHD filter using the set-based metrics. Regarding the detectors, we evaluate the performance when using (i) upper body detectors, (ii) face detectors or (iii) both. For each of these three choices, we also compare when adding color histogram descriptors or when not using them. From now on, U and F denote the use of upper-body detectors and face detectors respectively, while C denotes the use of color histograms.

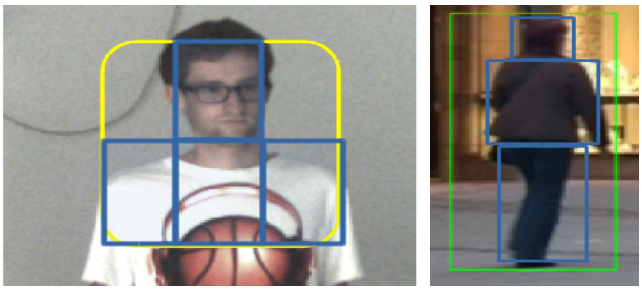
Results in Table 1 show that for the sequence CPD-2, while **Pr** and **MOTP** are higher when using upper-body detections U/UC, **Rc** and **MOTA** are higher when using face detections F/FC. One may think that the representation power of both detections may be complementary to each other. This is evidenced in the third row of Table 1, where both detectors are used and the performances are higher than in the first two rows, except for **Pr** and **MOTP** when using color. Regarding CPD-3, we clearly notice that the use of upper-body detections is much more advantageous than using the face detector. Importantly, even if the performance reported by the combination of the two detectors does not significantly outperform the ones reported when using only the upper-body detectors, it exhibits significant gains when compared to using only face detectors. The use of color seems to be advantageous in most of the cases, independently of the sequence and the detections used. Summarizing, while the performance of the method using only face detections or upper-body detections seems to be sequence-dependent, there is a clear advantage of using the feature combinations. Indeed,

[3] `http://motchallenge.net/`

(a) CPD-2 (b) CPD-3 (c) PETS09S2L1

(d) TUD-Stadtmitte (e) ParkingLot (f) TownCentre

Figure 3: Typical images extracted from the sequences used for tracking evaluation. Figures 3(a) and 3(b) are from the Cocktail-Party Dataset. Figures 3(c), 3(d), 3(e), 3(f) display sample images from PETS09S2L1, TUD-Stadtmitte, ParkingLot, and TownCentre which classically used in computer vision to evaluate multi-person tracking.



(a) (b)

Figure 4: Region splitting for computing the color histograms: Fig.4(a) shows an example with upper-body detection while Fig.4(b) shows an example of full body detection.

the combination seems to perform comparably to the best of the two detectors and much better to the worst. Therefore, the use of the combined detection appears to be the safest choice in the absence of any other information and therefore justifies developing a model able to handle observations coming from multiple detectors.

Table 2 reports a comparison of the proposed VEM model with the PHD filter for different features under the set metrics over the two sequences of the cocktail party dataset. We first observe that the behavior described from the results of Table 1 is also observed here, for a different group of measures and also for the PHD filter. Absolutely, while the use of the face or of the upper-body detections may be slightly more advantageous than the combination of detectors, this is sequence- and measure-dependent. However, the gain of the combination over the less reliable detector is very large, thus justifying the multiple-detector strategy when the applicative scenario allows for it and no other information about the sequence is available. The second observations is that the proposed VEM outperforms the PHD filter almost everywhere (i.e. except for CDP-3 with FU/FUC under the Hausdorff measure). This systematic trend demonstrates the potential of the proposed method from an experimental point of view. One possible explanation maybe that the variational tracker exploits additional information as it jointly estimates the target kinematic states together with their identities.

Figure 5 gives the histograms of the number of persons estimation absolute errors made by the variational tracking model. These results shows that for over the Cocktail Party Dataset, the number of people present in the visual scene for in a given time frame are in general correctly estimated. This shows that birth and the visibility processes play their role in creating tracks when new people enter the scene, and when they are occluded or leave the scene. More than 80% of the time, the correct

Sequence	Features	**Rc**	**Pr**	**MOTA**	**MOTP**
CPD-2	U/UC	53.3/70.7	94.9/99.4	46.6/64.3	80.8/85.8
	F/FC	89.8/90.1	94.6/94.6	75.7/76.0	76.6/76.7
	FU/FUC	93.1/95.2	95.3/96.2	88.3/80.0	76.5/82.9
CPD-3	U/UC	93.6/93.6	94.4/99.6	91.6/91.8	85.0/86.8
	F/FC	62.5/62.8	97.6/98.4	58.9/59.7	68.5/68.4
	FU/FUC	91.0/92.6	99.4/99.7	88.3/90.1	76.5/82.9

Table 1: Evaluation of the proposed multi-person tracking method with different features on the two sequences of the cocktail party dataset. All measures are in %.

Sequence	Method-Features	**Hausdorff**	**OMAT**	**OSPA**
CPD-2	VEM-U/UC	239.4/239.2	326.5/343.1	247.8/244.5
	PHD-U	276.6	435.3	567
	VEM-F/FC	116.3/115.5	96.3/96.1	110.9/108.0
	PHD-F	124	102	185.8
	VEM-FU/FUC	98.0/97.7	80.3/7	92.7/90.6
	PHD-FU	95	80	168
CPD-3	VEM-U/UC	56.0/56.2	44.4/44.2	54.7/54.1
	PHD-U	162.2	244.6	382.6
	VEM-F/FC	184.2/185.5	200.8/201.3436	203.3/205.0
	PHD-F	208	239.5	445.2
	VEM-FU/FUC	66.3/67.4	52.7/52.8	68.5/68.0
	PHD-FU	49	54.4	181

Table 2: Set metric based multi-person tracking performance measures of the proposed VEM and of the GMM PHD filter [44] on the the cocktail party dataset.

number of people is correctly estimated. It has to be noticed that errors are slightly higher for the sequence involving three person than for the sequence involving two persons.

To give a qualitative flavor to the tracking performance, Figure 9 gives sample results achieved by the proposed model (VEM-FUC) on CPD-3. These images show that the model is able to correctly initialize new tracks, identify occluded people as no longer visible, and recover their identities after occlusion. Tracking results are provided as supplementary material.

Figure 7 gives the estimated targets visibility probabilities (see Section 5.2) for sequence CPD-3 with sample tracking images given in Figure 6. The person visibility show that tracking for person 1 and 2 starts at the beginning of the sequence, and person 3 arrives at frame 600. Also, person 1 is occluded between frames 400 and 450 (see fourth image in the first row, and first image in the second row of Figure 6).

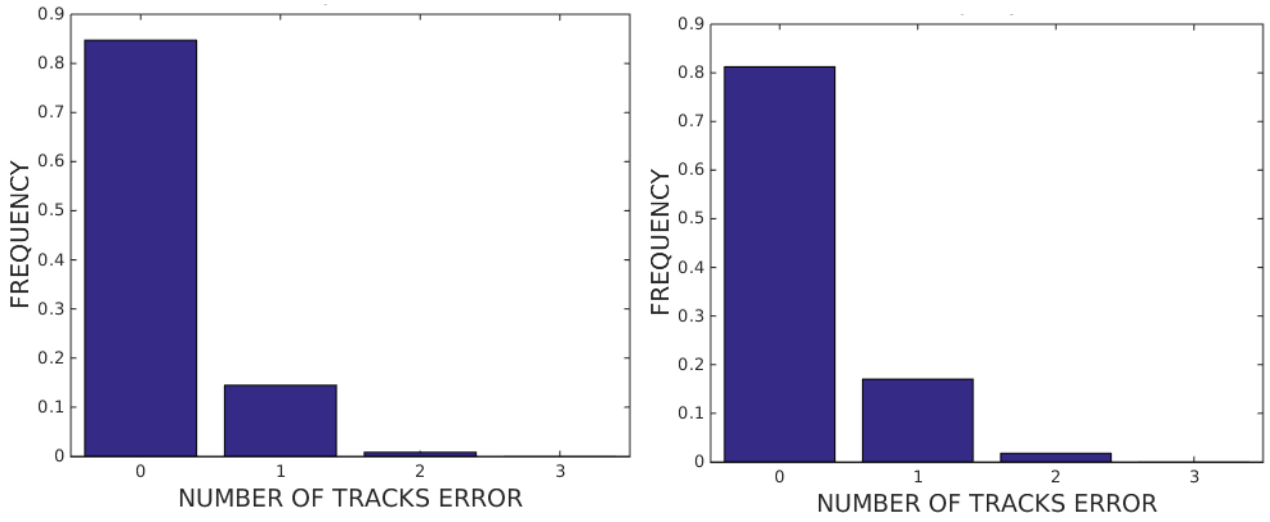
6.3. Evaluation on classical computer vision video sequences

In this tracking situation, we model a single person's kinematic state as the full body bounding box and its velocity. In this case, the observation operator $\mathbf{P}$ simply removes the velocity information, keeping only the bounding box' position and size. The appearance observations are the concatenation of the joint HS histograms of the head, torso and legs areas (see Figure 4(b)).

We evaluate our model using only body localization observations (B) and jointly using body localization and color appearance observations (BC). Table 3 compare the proposed variational model to the PHD filter using set based distance performance metrics. As for the cocktail party dataset, in general, these results show that the variational tracker outperforms the PHD filter.

In addition, we also compare the proposed model to two tracking models, proposed by Milan *et al* in [18] and by Bae and Yoon in [31]. Importantly, the direct comparison of our model to these two state-of-the-art methods must be done with care. Indeed, while the proposed VEM uses only causal (past) information, these two methods use both past and future detections. In other words, while ours is a *filtering*, [18, 31] are *smoothing* methods. Therefore, we expect these two models to outperform the proposed one. However, the main prominent advantage of filtering methods over smoothing methods, and therefore of the proposed VEM over these two methods, is that while smoothing methods are inherently unsuitable for on-line processing, filtering methods are naturally appropriate for on-line task, since they only use causal information.

Table 4 reports the performance of these methods on four sequences classically used in computer vision to evaluate multi-target trackers. In this table, results over TUD-Stadmitte show similar performances for our model using or not appearance information. Therefore, color information is not very informative in this sequence. In PETS09-S2-L1, our model using color achieves better MOTA measure, precision, and recall, showing



(a) CPD-2 (b) CPD-3

Figure 5: Histogram of absolute errors about the estimation of the number of people present in the visual scene over the Cocktail Party Dataset.

Figure 6: Sample tracking results on CPD-3. The green bounding boxes represent the face detections and the yellow bounding boxes represent the upper body detections. Importantly, the red bounding boxes display the tracking results.

the benefit of integrating color into the model. As expected, Milan *et al* and Bae and Yoon, outperform the proposed model. However, the non-causal nature of their method makes them unsuitable for on-line tracking tasks, where the observations must be processed when received, and not before.

Figure 8 gives the histograms of the errors about the number of people present in the visual scene for the four sequences ParkingLot, TownCentre, PETS09-S2L1, TUD-Stadtmitte. These results show that, the four sequences are more challenging than the Cocktail Party Dataset (see figure 5). Among the four video sequences, TUD-Stadtmitte is the one where variational tracking model is making the estimated number of people is the less consistent. This can be explained by the quality of the observations (detections) over this sequence. For the PETS, and the ParkingLot dataset which involve about 15 persons, about 70% of the time the proposed tracking model is estimating the number of people in the scene with an error below 2 persons. For the TownCentre sequence which involves 231 persons over 4500 frames, over 70% of the time, the error made by the variational tracker is below 7 persons. This shows that, even in challenging situations involving occlusions due to crowd, the birth and the visibility process play their role.

Figure 9 presents sample results for the PET09-S2L1 sequence. In addition, videos presenting the results on the second dataset are provided as supplementary material. These results show temporally consistent tracks. Occasionally, person identity switches may occur when two people cross. Remarkably, because the proposed tracking model is allowed to reuse the identity of persons visible in the past, people re-entering the scene after having left, will be recognized the the previously used track will be awaken.

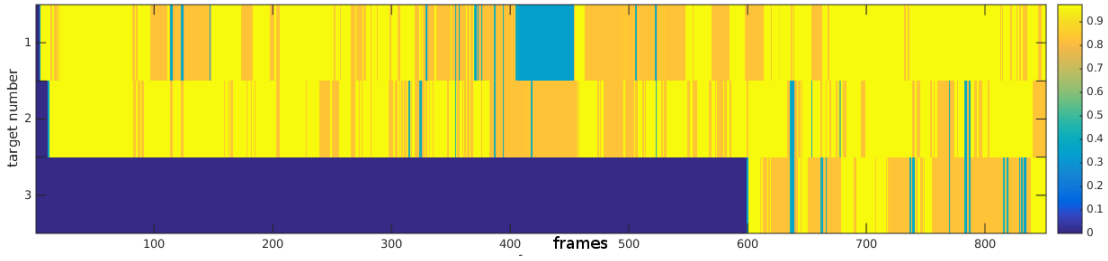


Figure 7: Estimated visibility probabilities for tracked persons in sequence CPD-3. Every row displays the corresponding targets visibility probabilities for every time frame. Yellow color represents very high probability (close to 1), and blue color represents very low probabilities.

Sequence	Method-Features	**Hausdorff**	**OMAT**	**OSPA**
TUD-Stadtmitte	VEM-B/BC	150.4/125.9	197.5/184.9	483.2/482.4
	PHD-B	184.7	119	676
PETS09S2L1	VEM-B/BC	52.1/50.9	72.6/40.8	117.0/110.1
	PHD-B	70	44	163
TownCentre	VEM-B/BC	420./391.2	205.4/177.5	350.0 /335.2
	PHD-B	430.5	173.8	364.9
ParkingLot	VEM-B/BC	95.0/90.5	87.9/83.9	210.8/203.4
	PHD-B	169	94.0	415

Table 3: Set metric based multi-person tracking Performance measures on the sequences the four sequences PETS09S2L1, TownCentre, ParkingLot,and TUD-Stadtmitte.

7. Conclusions

We presented an on-line variational Bayesian model to track a time-varying number of persons from cluttered multiple visual observations. Up to our knowledge, this is the first variational Bayesian model for tracking multiple persons, or more generally, multiple targets. We proposed birth and visibility processes to handle persons that are entering and leaving the visual field. The proposed model is evaluated with two datasets showing competitive results with respect to state of the art multi-person tracking models. Remarkably, even if in the conducted experiments we model the visual appearance with color histograms, our framework is versatile enough to accommodate other visual cues such as texture, feature descriptors or motion cues.

In the future we plan to consider the integration of more sophisticated birth processes than the one considered in this paper, e.g. [46]. We also plan to extend the visual tracker to incorporate auditory cues. For this purpose, we plan to jointly track the kinematic states and the speaking status (active/passive) of each tracked person. The framework proposed in this paper allows to exploit audio features, e.g. voice activity detection and audio-source localization as observations. When using audio information, robust voice descriptors (the acoustic equivalent of visual appearance) and their blending with the tracking model will be investigated. We also plan to extend the proposed formalism to a moving camera such that its kinematic state is tracked as well. This case is of particular interest in applications such as pedestrian tracking for self-driving cars or for human-robot interaction.

Appendix A. Derivation of the Variational Formulation

Appendix A.1. Filtering Distribution Approximation

The goal of this section is to derive an approximation of the hidden-state filtering distribution $p(\mathbf{Z}_t, \mathbf{X}_t|\mathbf{o}_{1:t}, \mathbf{e}_{1:t})$, given the variational approximating distribution $q(\mathbf{Z}_{t-1}, \mathbf{X}_{t-1})$ at $t-1$. Using Bayes rule, the filtering distribution can be written as

$$p(\mathbf{Z}_t, \mathbf{X}_t|\mathbf{o}_{1:t}, \mathbf{e}_{1:t}) = \frac{p(\mathbf{o}_t|\mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t)p(\mathbf{Z}_t, \mathbf{X}_t|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})}{p(\mathbf{o}_t|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})}. \quad \text{(A.1)}$$

It is composed of three terms, the likelihood $p(\mathbf{o}_t|\mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t)$, the predictive distribution $p(\mathbf{Z}_t, \mathbf{X}_t|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})$, and the normalization factor $p(\mathbf{o}_t|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})$ which is independent of the hidden variables. The likelihood can be expanded as:

$$p(\mathbf{o}_t|\mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t) = \prod_{i=1}^{I} \prod_{k \leq K_t^i} \prod_{n=0}^{N} p(\mathbf{o}_{tk}|Z_{tk} = n, \mathbf{X}_t, \mathbf{e}_t)^{\delta_n(Z_{tk}^i)} \quad \text{(A.2)}$$

where δ_n is the Dirac delta function, and $p(\mathbf{o}_{tk}|Z_{tk} = n, \mathbf{X}_t, \mathbf{e}_t)$ is the individual observation likelihood defined in (5) and (6).

The predictive distribution factorizes as

$$p(\mathbf{Z}_t, \mathbf{X}_t|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t}) = p(\mathbf{Z}_t|\mathbf{e}_t)p(\mathbf{X}_t|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t}).$$

Exploiting its multinomial nature, the assignment variable distribution $p(\mathbf{Z}_t|\mathbf{e}_t)$ can be fully expanded as:

$$p(\mathbf{Z}_t|\mathbf{e}_t) = \prod_{i=1}^{I} \prod_{k \leq K_t^i} \prod_{n=0}^{N} p(Z_{tk}^i = n|\mathbf{e}_t)^{\delta_n(Z_{tk}^i)}. \quad \text{(A.3)}$$

Sequence	Method	Rc	Pr	MOTA	MOTP
TUD-Stadmitte	VEM-B/BC	72.2/70.9	81.7/82.5	54.8/53.5	65.4/65.1
	[18]	84.7	86.7	71.5	65.5
PETS09-S2L1	VEM-B/BC	90.1/90.2	86.2/87.6	74.9/76.7	71.8/71.8
	[18]	92.4	98.4	90.6	80.2
	[31]	-	-	83	69.5
TownCentre	VEM-B/BC	88.1/90.1	71.5/72.7	72.7/70.9	74.9/76.1
ParkingLot	VEM-B/BC	80.3/78.3	85.2/87.5	73.1/74	70.8/71.7

Table 4: Performance measures on the sequences of the second dataset. Comparison with [18, 31] must be done with care since both are smoothing methods and therefore use more information than the proposed VEM.

Using the motion state dynamics definition $p(\mathbf{x}_{tn}|\mathbf{x}_{t-1n}, \mathbf{e}_{tn})$ the previous time motion state filtering distribution variational approximation $q(\mathbf{x}_{t-1n}|\mathbf{e}_{t-1}) = p(\mathbf{x}_{t-1n}|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t-1})$ defined in (20), motion state predictive distribution $p(\mathbf{X}_t = \mathbf{x}_t|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})$ can approximated by

$$\begin{aligned}
&p(\mathbf{X}_t = \mathbf{x}_t|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t}) \\
&= \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{e}_t) p(\mathbf{x}_{t-1}|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t-1}) d\mathbf{x}_{t-1} \\
&= \int \left(\prod_{n=1}^{N} p(\mathbf{x}_{tn}|\mathbf{x}_{t-1n}, \mathbf{e}_{tn}) \right) p(\mathbf{x}_{t-1}|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t-1}) d\mathbf{x}_{t-1} \\
&\approx \int \prod_{n=1}^{N} p(\mathbf{x}_{tn}|\mathbf{x}_{t-1n}, \mathbf{e}_{tn}) q(\mathbf{x}_{t-1n}|\mathbf{e}_{t-1n}) d\mathbf{x}_{t-1,1} ... d\mathbf{x}_{t-1,n} \\
&\approx \prod_{n=1}^{N} u(\mathbf{x}_{tn})^{1-e_{tn}} g(\mathbf{x}_{tn}, \mathbf{D}\boldsymbol{\mu}_{t-1,n}, \mathbf{D}\boldsymbol{\Gamma}_{tn}\mathbf{D}^\top + \boldsymbol{\Lambda}_n)^{e_{tn}} \qquad \text{(A.4)}
\end{aligned}$$

where during the derivation, the filtering distribution of the kinematic state at time $t-1$ is replaced by its variational approximation $p(\mathbf{x}_{t-1}|\mathbf{o}_{1:t-1}, \mathbf{e}_{1:t-1}) = \prod_{n=1}^{N} q(\mathbf{x}_{t-1n}|\mathbf{e}_{t-1n})$.

Equations (A.2), (A.3), and (A.4) define the numerator of the tracking filtering distribution (A.1). The logarithm of this filtering distribution is used by the proposed variational EM algorithm.

Appendix A.2. Derivation of the E-Z-Step

The E-Z-step corresponds to the estimation of $q(Z^i_{tk}|\mathbf{e}_t)$ given by (14) which, from the log of the filtering distribution, can be written as:

$$\begin{aligned}
&\log q(Z^i_{tk}|\mathbf{e}_t) = \\
&\sum_{n=0}^{N} \delta_n(Z^i_{tk}) \mathbf{E}_{q(\mathbf{X}_{tn}|\mathbf{e}_t)}[\log \left(p(\mathbf{y}^i_{tk}, \mathbf{h}^i_{tk}|Z^i_{tk} = n, \mathbf{X}_t, \mathbf{e}_t) p(Z^i_{tk} = n|\mathbf{e}_t) \right)] + C, \qquad \text{(A.5)}
\end{aligned}$$

where C gathers terms that are constant with respect to the variable of interest, Z^i_{tk} in this case. By substituting $p(\mathbf{y}^i_{tk}, \mathbf{h}^i_{tk}|Z^i_{tk} = n, \mathbf{X}_t, \mathbf{e}_t)$, and $p(Z^i_{tk} = n|\mathbf{e}_t)$ with their expressions (5), (6), and (8), by introducing the notations

$$\begin{aligned}
\epsilon^i_{tk0} &= u(\mathbf{y}^i_{tk}) u(\mathbf{h}^i_{tk}) \\
\epsilon^i_{tkn} &= g(\mathbf{y}^i_{tk}, \mathbf{P}\boldsymbol{\mu}_{tn}, \boldsymbol{\Sigma}^i) \exp(-\frac{1}{2} \mathrm{Tr}(\mathbf{P}^\top {\boldsymbol{\Sigma}^i}^{-1} \mathbf{P}\boldsymbol{\Gamma}_{tn})) b(\mathbf{h}^i_{tk}, \mathbf{h}_n)
\end{aligned}$$

and after some algebraic derivations, the distribution of interest can be written as the following multinomial distribution

$$q(Z^i_{tk} = n|\mathbf{e}_t) = \alpha^i_{tkn} = \frac{e_{tn} \epsilon^i_{tkn}}{\sum_{m=0}^{N} e_{tm} \epsilon^i_{tkm}} \qquad \text{(A.6)}$$

Appendix A.3. Derivation of the E-X-Step

The E-step for the motion state variables consists in the estimation of $q(\mathbf{X}_{tn}|e_{tn})$ using relation $\log q(\mathbf{X}_{tn}|e_{tn}) = \mathbf{E}_{q(Z_t, \mathbf{X}_t/\mathbf{X}_{tn}|\mathbf{e}_t)}[\log p(\mathbf{Z}_t, \mathbf{X}_t|\mathbf{o}_{1:t}, \mathbf{e}_{1:t})]$ which can be expanded as

$$\begin{aligned}
\log q(\mathbf{X}_{tn}|\mathbf{e}_t) = \sum_{i=1}^{I} \sum_{k=0}^{K^i_t} \mathbf{E}_{q(Z^i_{tk}|\mathbf{e}_t)}[\delta_n(Z^i_{tk})] \log g(\mathbf{y}^i_{tk}; \mathbf{P}\mathbf{X}_{tn}, \boldsymbol{\Sigma}^i)^{e_{tn}} \\
+ \log(u(\mathbf{X}_{tn})^{1-e_{tn}} g(\mathbf{X}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1n}, \mathbf{D}\boldsymbol{\Gamma}_{tn}\mathbf{D}^\top + \boldsymbol{\Lambda}_n)^{e_{tn}}) + C,
\end{aligned}$$

where, as above, C gathers constant terms. After some algebraic derivation one obtains $q(\mathbf{X}_{tn}|e_{tn}) = u(\mathbf{X}_{tn})^{1-e_{tn}} g(\mathbf{X}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn})^{e_{tn}}$ where the mean and covariance of the Gaussian distribution are given by (21) and by (22).

Acknowledgments

Support from the ERC Advanced Grant VHIA #34113, from the EU-FP7 STREP project EARS #609465, and from MIUR Active Aging at Home #CTN01 00128 is greatly acknowledged.

References

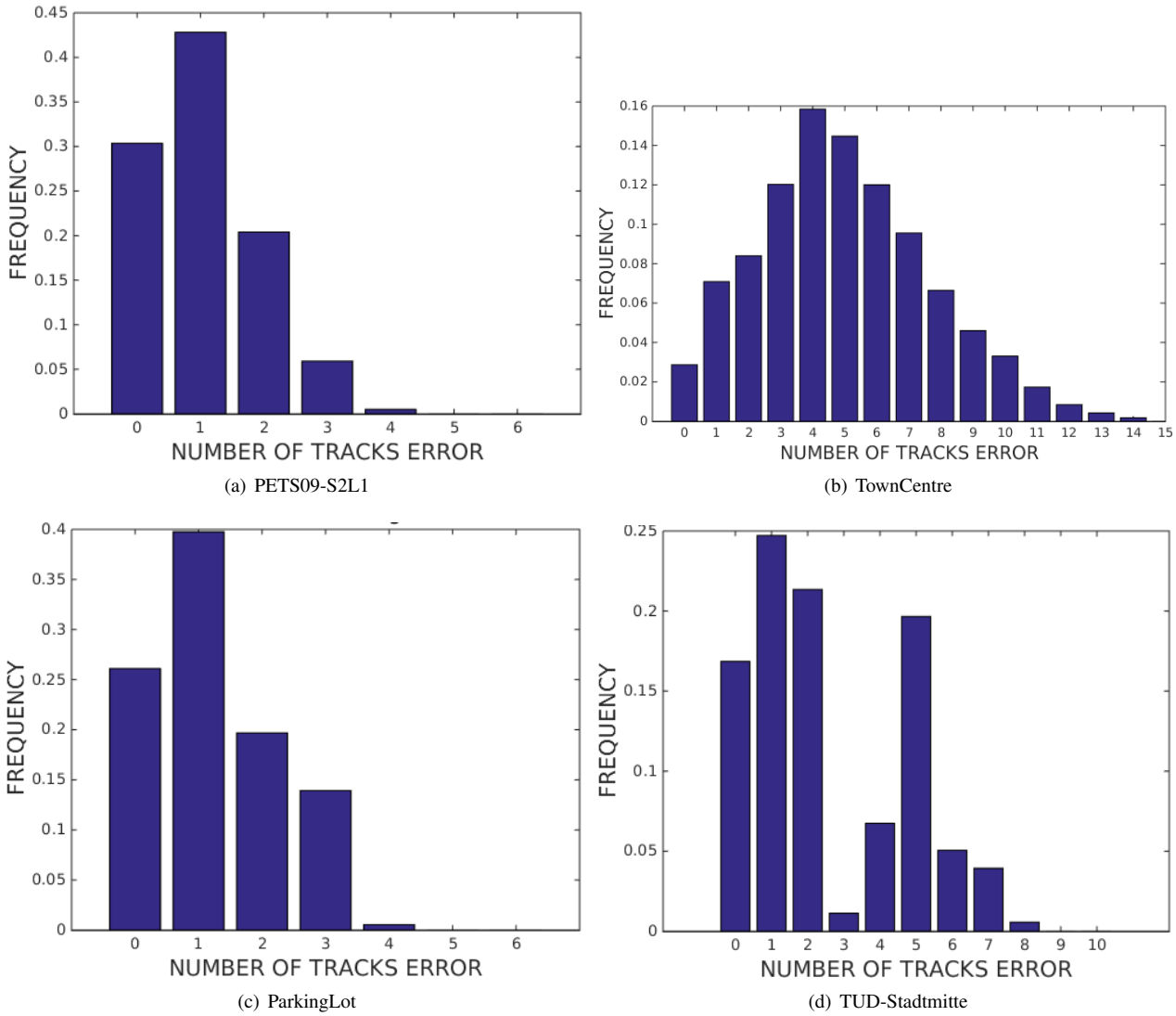
[1] W. Luo, J. Xing, X. Zhang, W. Zhao, T.-K. Kim, Multiple object tracking: a review, arXiv:1409.761 (2015).

[2] M. Andriluka, S. Roth, B. Scheile, People-tracking-by-detection and people-detection-by-tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008, pp. 1–8.

[3] P. J. Green, Trans-dimensional Markov chain Monte Carlo, in: Oxford Statistical Science Series, 2003, pp. 179–198.

[4] Z. Khan, T. Balch, F. Dellaert, An MCMC-based particle filter for tracking multiple interacting targets, in: European Conference on Computer Vision, Prague, Czech Republic, 2004, pp. 279–290.

[5] K. Smith, D. Gatica-Perez, J.-M. Odobez, Using particles to track varying numbers of interacting people, in: IEEE Computer Vision and Pattern Recognition, San Diego, USA, 2005, pp. 962–969.



(a) PETS09-S2L1

(b) TownCentre

(c) ParkingLot

(d) TUD-Stadtmitte

Figure 8: Histogram of errors about the estimation of the number of people present in the visual scene over ParkingLot, TownCentre, PETS09-S2L1, TUD-Stadtmitte.

[6] M. Yang, Y. Liu, L. Wen, Z. You, A probabilistic framework for multi-target tracking with mutual occlusions, IEEE Confenrence on Computer Vision and Pattern Recognition (2014) 1298 – 1305.

[7] P. J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (4) (1995) 711–732.

[8] R. P. Mahler, Multisource multitarget filtering: a unified approach, in: Aerospace/Defense Sensing and Controls, International Society for Optics and Photonics, 1998, pp. 296–307.

[9] R. P. S. Mahler, Statistics 101 for multisensor, multitarget data fusion, IEEE Aerospace and Electronic Systems Magazine 19 (1) (2004) 53–64.

[10] R. P. S. Mahler, Statistics 102 for multisensor multitarget data fusion, IEEE Selected Topics on Signal Processing 19 (1) (2013) 53–64.

[11] R. P. S. Mahler, A theoretical foundation for the Stein-Winter" probability hypothesis density (PHD)" multitarget tracking approach, Tech. rep. (2000).

[12] H. Sidenbladh, Multi-target particle filtering for the probability hypothesis density, in: IEEE International Conference on Information Fusion, Tokyo, Japan, 2003, pp. 800–806.

[13] D. Clark, J. Bell, Convergence results for the particle PHD filter, IEEE Transactions on Signal Processing 54 (7) (2006) 2652–2661.

[14] B.-N. Vo, S. Singh, A. Doucet, Random finite sets and sequential monte carlo methods in multi-target tracking, in: IEEE International Radar Conference, Huntsville, USA, 2003, pp. 486–491.

[15] W. K. Ma, B. N. Vo, S. S. Singh, Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach, IEEE Transactions on Signal Processing 54 (9) (2006) 3291–3304.

[16] E. Maggio, M. Taj, A. Cavallaro, Efficient multitarget visual tracking using random finite sets, IEEE Transactions on Circuits and Systems for Video Technology 18 (8) (2008) 1016–1027.

[17] B. Yang, R. Nevatia, An online learned CRF model for multi-target tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 2034–2041.

[18] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for multitarget tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (1) (2014) 58–72.

[19] A. Heili, A. Lopez-Mendez, J.-M. Odobez, Exploiting long-term connectivity and visual motion in CRF-based multi-person tracking, IEEE Transactions on Image Processing 23 (7) (2014) 3040–3056.

[20] Y. Bar-Shalom, F. Daum, J. Huang, The probabilistic data association filter: estimation in the presence of measurement origin and uncertainty, IEEE Control System Magazine 29 (6) (2009) 82–100.

[21] J. Vermaak, N. Lawrence, P. Perez, Variational inference for visual tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, Madison, USA, 2003, pp. 773–780.

[22] V. Smidl, A. Quinn, The Variational Bayes Method in Signal Processing, Springer, 2006.

Figure 9: Tracking results on PETS09-S2L1. Green boxes represent observations and red bounding boxes represent tracking outputs associated with person identities. Green and red bounding boxes may overlap.

[23] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 2007.
[24] A. Yilmaz, O. Javed, M. Shah, Object tracking: A survey, in: ACM Computing Surveys, 2006, p. 13.
[25] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1627–1645.
[26] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: IEEE Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 2879–2886.
[27] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619.
[28] S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian bayesian tracking, IEEE Transactions on Signal Processing 50 (2) (2002) 174–188.
[29] S. Sarka, A. Vehtari, J. Lampinen, Rao-Blackwellized Monte Carlo data association for multiple target tracking, in: IEEE International Conference on Information Fusion, Stockholm, Sweden, 2004, pp. 583–590.
[30] Y. Yan, A. Kostin, W. Christmas, J. Kittler, A novel data association algorithm for object tracking in clutter with application to tennis video analysis, in: IEEE Computer Vision and Pattern Recognition, New York, USA, 2006, pp. 634–641.
[31] S.-W. Bae, K.-J. Yoon, Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning, in: IEEE Computer Vision and Pattern Recognition, Columbus, USA, 2014, pp. 1218–1225.
[32] H. Pirsiavash, D. Ramanan, C. C. Fowlkes, Globally-optimal greedy algorithms for tracking a variable number of objects, IEEE Conference on Computer Vision and Pattern Recognition (2011) 1201–1208.
[33] M. Isard, J. MacCormick, Bramble: A bayesian multiple-blob tracker, in: IEEE International Conference on Computer Vision, British Columbia, Canada, 2001, pp. 34–41.
[34] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, D. G. Lowe, A boosted particle filter: Multitarget detection and tracking, in: European Conference on Computer Vision, Prague, Czech Republic, 2004, pp. 28–39.
[35] S. Sarka, A. Vehtari, J. Lampinen, Rao-Blackwellized particle filter for multiple target tracking, in: IEEE International Conference Information Fusion, Québec ,Canada, 2007, pp. 2–15.
[36] A. R. Zamir, A. Dehghan, M. Shah, Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs, IEEE Conference on Computer Vision and Pattern Recognition (2012) 343–356.
[37] W. Longyin, W. Li, J. Yan, Z. Lei, D. Yi, S. Z. Li, Multiple target tracking based on undirected hierarchical relation hypergraph, IEEE Conference on Computer Vision and Pattern Recognition (2014) 1282–1289.
[38] P. Perez, C. Hue, J. Vermaak, M. Gangnet, Color-based probabilistic tracking, in: European Conference on Computer Vision, Copenhagen, Denmark, 2002, pp. 661–675.
[39] L. Leal-Taixe, A. Milan, I. Reid, S. Roth, K. Schindler, MOT Challenge 2015: Towards a benchmark for multi-target tracking, arXiv:1504.01942 (2015).
[40] B. Ristic, B.-N. Vo, D. Clark, Performance evaluation of multi-target tracking using the OSPA metric, in: IEEE International Conference on Information Fusion, Edinburgh, UK, 2010, pp. 1–7.
[41] K. Smith, D. Gatica-Perez, J.-M. Odobez, S. Ba, Evaluating multi-object tracking, in: IEEE CVPR Workshop on Empirical Evaluation Methods in Computer Vision, San Diego, USA, 2005, pp. 36–36.
[42] R. Stiefelhagen, K. Bernardin, R. Bowers, J. S. Garofolo, D. Mostefa, P. Soundararajan, CLEAR 2006 evaluation, in: First International Workshop on Classification of Events and Relationship, CLEAR 2006, Springer, 2005.
[43] W. Longyin, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, S. Lyu, DETRAC filter multiple target tracker: A new benchmark and protocol for multi-object tracking, arXiv:1511.04136.
[44] D. E. Clark, K. Panta, B. N. Vo, The GM-PHD filter multiple target tracker, IEEE International Conference In Information Fusion (2006) 1–8.
[45] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, D. Ferguson, Real-time pedestrian detection with deep network cascades, British Machine Vision Conference.
[46] R. Streit, Birth and death in multitarget tracking filters, in: IEEE Workshop on Sensor Data Fusion: Trends, Solutions, Applications, 2013, pp. 1–6.