

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

Sileye Ba<sup>a,\*</sup>, Xavier Alameda-Pineda<sup>a,b</sup>, Alessio Xompero<sup>a</sup>, Radu Horaud<sup>a</sup>

<sup>a</sup>*INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France*

<sup>b</sup>*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.

---

\*Corresponding author

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

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

---

## 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

36 tracking a varying number of objects. The main drawback is that the states  
37 are inferred by means of a reversible jump Markov chain Monte Carlo sam-  
38 pling, which is computationally expensive [7]. Secondly, a random finite set  
39 multi-target tracking formulation was proposed [8, 9, 10]. Initially used for  
40 radar applications [8], in this framework the targets are modeled as real-  
41 izations of a random finite set which is composed of an unknown number of  
42 elements. Because an exact solution to this model is computationally inten-  
43 sive, an approximation known as the probability hypothesis density (PHD)  
44 filter was proposed [11]. Further sampling-based approximations of random  
45 set based filters were subsequently proposed, e.g. [12, 13, 14]. These were  
46 exploited in [15] for tracking a time-varying number of active speakers using  
47 auditory cues and in [16] for multi-target tracking using visual observations.  
48 Thirdly, conditional random fields (CRF) were also chosen to address multi-  
49 target tracking [17, 18, 19]. In this case, tracking is casted into an energy  
50 minimization problem. In another line of research, in radar tracking, other  
51 popular multi-targets tracking model are joint probabilistic data association  
52 (JPDA), and multiple hypothesis filters [20].

53 In this paper we propose an on-line variational Bayesian framework for  
54 tracking an unknown and varying number of persons. Although variational  
55 model are very popular in machine learning, their use in computer vision for  
56 object tracking has been limited to tracking situation involving a fixed num-  
57 ber of targets [21]. Variational Bayes methods approximate the joint a poste-  
58 riori distribution of the latent variables by a separable distribution [22, 23].  
59 In an on-line tracking scenario, where only causal (past) observations can  
60 be used, this translates into approximating the filtering distribution. This  
61 is in strong contrast with off-line trackers that use both past and future  
62 observations. The proposed tracking algorithm is therefore modeling the a  
63 posteriori distribution of the hidden states given all past observations. Im-  
64 portantly, the proposed framework leads to closed-form expressions for the  
65 posterior distributions of the hidden variables and for the model parameters,  
66 thus yielding an intrinsically efficient filtering procedure implemented via an  
67 variational EM (VEM) algorithm. In addition, a *clutter target* is defined so  
68 that spurious observations, namely detector failures, are associated to this  
69 target and do not contaminate the filtering process. Furthermore, our for-  
70 malism allows to integrate in a principled way heterogeneous observations  
71 coming from various detectors, e.g. face, upper-body, silhouette, etc. Re-  
72 markably, objects that come in and out of the field of view, namely object  
73 appearance and disappearance, are handled by object birth and visibility

74 processes. In details, we replace the classical death process by a visibility  
 75 process which allows to put asleep tracks associated with persons that are  
 76 no longer visible. The main advantage is that these tracks can be awoken  
 77 as soon as new observations match their appearance with high confidence.  
 78 Summarizing, the paper contributions are:

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

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

## 100 2. Related Work

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

107 into single-object tracking, fixed-number multi-object tracking, and varying-  
108 number multi-object tracking.

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

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

133     In most practical applications, the number of objects to be tracked, is  
134 not only unknown, but it also varies over time. Importantly, tracking a  
135 time-varying number of objects requires an efficient mechanism to add new  
136 objects entering the field of view, and to remove objects that moved away.  
137 In a probabilistic setting, these mechanisms are based on birth and death  
138 processes. Efficient multi-object algorithms have to be developed within  
139 principled methodologies allowing to handle hidden states of varying dimen-  
140 sionality. In computer vision, the most popular methods are based on condi-  
141 tional random fields [31, 18, 19, 32], on random finite sets [10, 15, 16] or on  
142 the trans-dimensional Markov chain [3, 4, 5, 6]. [6] presents an interesting  
143 approach where occlusion state of a tracked person is explicitly modeled in  
144 the tracked state and used for observation likelihood computation. Less pop-

145 ular but successful methodologies include the Bayesian multiple blob tracker  
 146 of [33], the boosted particle filter for multi-target tracking of [34] and the  
 147 Rao-Blackwellized filter for multiple objects tracking [35], graph based rep-  
 148 resentation for multi-object tracking [36, 37]. It has to be noticed in other  
 149 communities, such as radar tracking, multi-object tracking has been deeply  
 150 investigated. Many models have been proposed such as the probabilistic data  
 151 association filter (PDAF), the joint PDAF, multiple hypothesis tracking [20].  
 152 However, the differences between multi-object tracking in radar and in com-  
 153 puter vision are mainly two. On the one hand, most tracking method for  
 154 radar consider point-wise objects, modeling a punctual latent state, whereas  
 155 in computer vision objects are represented using bounding boxes in addition  
 156 to the punctual coordinates. On the other hand, computer vision applica-  
 157 tions benefit from the use of visual appearance, which is mainly used for  
 158 object identification [38].

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

### 178 3. Variational Bayesian Multiple-Person Tracking

#### 179 3.1. Notations and Definitions

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

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

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

201 In order to ease the challenging task of tracking multiple persons with a  
 202 single static camera, we assume the existence of  $I$  detectors, each of them  
 203 providing  $K_t^i$  localization observations at each time  $t$ , with  $i \in [1 \dots I]$ . Fig. 1  
 204 provides examples of face and upper-body detections (see Fig. 1(a)) and  
 205 of full-body detections (see Fig. 1(b)). The  $k$ -th localization observation  
 206 gathered by the  $i$ -th detector at time  $t$  is denoted by  $\mathbf{y}_{tk}^i \in \mathbb{R}^4$ , and represents  
 207 the location (2D position, width, height) of a person in the image. The  
 208 set of observations provided by detector  $i$  at time  $t$  is denoted by  $\mathbf{y}_t^i =$   
 209  $\{\mathbf{y}_{tk}^i\}_{k=1}^{K_t^i}$ , and the observations provided by all the detectors at time  $t$  is  
 210 denoted by  $\mathbf{y}_t = \{\mathbf{y}_t^i\}_{i=1}^I$ . Associated to each localization detection  $\mathbf{y}_{tk}^i$ , there  
 211 is a photometric description of the person’s appearance, denoted by  $\mathbf{h}_{tk}^i$ .  
 212 This photometric observation is extracted from the bounding box of  $\mathbf{y}_{tk}^i$ .  
 213 Altogether, the localization and photometric observations constitute the raw  
 214 observations  $\mathbf{o}_{tk}^i = (\mathbf{y}_{tk}^i, \mathbf{h}_{tk}^i)$  used by our tracker. Analogous definitions to  $\mathbf{y}_t^i$



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.

215 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  $\mathbf{o}_t = \{\mathbf{o}_t^i\}_{i=1}^I$ .  
 216 Importantly, when we write the probability of a set of random variables,  
 217 we refer to the joint probabilities of all random variables in that set. For  
 218 instance:  $p(\mathbf{o}_t^i) = p(\mathbf{o}_{t1}^i, \dots, \mathbf{o}_{tK_t^i}^i)$ .

219 We also define an observation-to-person assignment (hidden) variable  $Z_{tk}^i$   
 220 associated with each observation  $\mathbf{o}_{tk}^i$ . Formally,  $Z_{tk}^i$  is a categorical variable  
 221 taking values in the set  $\{1 \dots N\}$ :  $Z_{tk}^i = n$  means that  $\mathbf{o}_{tk}^i$  is associated to  
 222 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  
 223 assignment variables can be easily used to handle false detections. Indeed, it  
 224 is common that a detection corresponds to some clutter instead of a person.  
 225 We cope with these false detections by defining a *clutter* target. In practice,  
 226 the index  $n = 0$  is assigned to this clutter target, which is always visible,  
 227 i.e.  $e_{t0} = 1$  for all  $t$ . Hence, the set of possible values for  $Z_{tk}^i$  is extended to  
 228  $\{0\} \cup \{1 \dots N\}$ , and  $Z_{tk}^i = 0$  means that observation  $\mathbf{o}_{tk}^i$  has been generated  
 229 by clutter and not by a person. The practical consequence of adding a clutter  
 230 track is that the observations assigned to it play no role in the estimation of  
 231 the parameters of the other tracks, thus leading to estimation rules inherently  
 232 robust to outliers.

### 233 3.2. The Proposed Bayesian Multi-Person Tracking Model

234 The on-line multi-person tracking problem is cast into the estimation of  
 235 the filtering distribution of the hidden variables given the causal observations

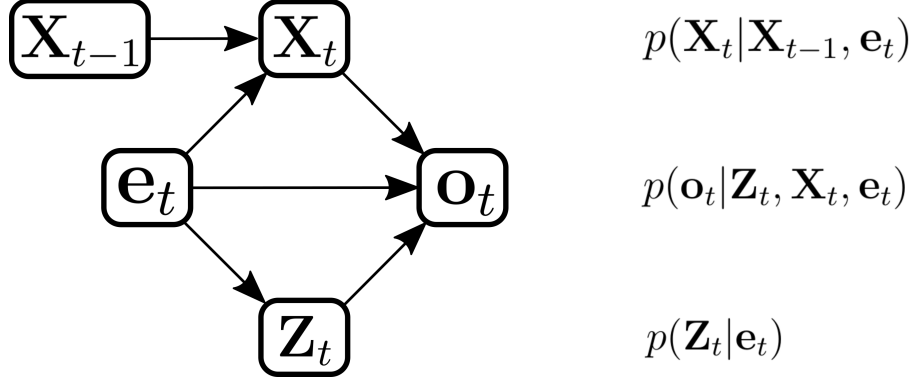


Figure 2: Graphical representation of the proposed multi-target tracking probabilistic model.

236  $p(\mathbf{Z}_t, \mathbf{X}_t | \mathbf{o}_{1:t}, \mathbf{e}_{1:t})$ , where  $\mathbf{o}_{1:t} = \{\mathbf{o}_1, \dots, \mathbf{o}_t\}$ . The filtering distribution can  
 237 be rewritten as:

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

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

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

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

### 248 3.2.1. The Observation Model

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

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

251 In addition, we make the reasonable assumption that, while localization ob-  
 252 servations depend both on the assignment variable and kinematic state, the  
 253 appearance observations only depend on the assignment variable, that is the  
 254 person identity, but not on his/her kinematic state. We also assume the lo-  
 255 calization and appearance observations to be independent given the hidden  
 256 variables. Consequently, the observation likelihood of a single joint observa-  
 257 tion can be factorized as:

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

258 The localization observation model is defined depending on whether the ob-  
 259 servation is generated by clutter or by a person:

- 260 • If the observation is generated from clutter, namely  $Z_{tk}^i = 0$ , the vari-  
 261 able  $\mathbf{y}_{tk}^i$  follows an uniform distribution with probability density func-  
 262 tion  $u(\mathbf{y}_{tk}^i)$ ;
- 263 • If the observation is generated by person  $n$ , namely  $Z_{tk}^i = n$ , the vari-  
 264 able  $\mathbf{y}_{tk}^i$  follows a Gaussian distribution with mean  $\mathbf{P}^i \mathbf{X}_{tn}$  and covari-  
 265 ance  $\Sigma^i$ :  $\mathbf{y}_{tk}^i \sim g(\mathbf{y}_{tk}^i; \mathbf{P}^i \mathbf{X}_{tn}, \Sigma^i)$

266 The linear operator  $\mathbf{P}^i$  maps the kinematic state vectors onto the  $i$ -th space of  
 267 observations. For example, when  $\mathbf{X}_{tn}$  represents the upper-body kinematic  
 268 state (upper-body localization and velocity) and  $\mathbf{y}_{tk}^i$  represents the upper-  
 269 body localization observation,  $\mathbf{P}^i$  is a projection which, when applied to a  
 270 state vector, only retains the localization components of the state vector.  
 271 When  $\mathbf{y}_{tk}^i$  is a face localization observation, the operator  $\mathbf{P}^i$  is a composition  
 272 of the previous projection, and an affine transformation mapping an upper-  
 273 body bounding-box onto its corresponding face bounding-box. Finally, the  
 274 full observation model is compactly defined by

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

275 where  $\delta_{ij}$  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 clut-  
 ter, the appearance observation follows a uniform distribution with density

function  $u(\mathbf{h}_{tk}^i)$ . When the observation is generated by person  $n$ , the appearance observation follows a Bhattacharya distribution with density defined as

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

where  $\lambda$  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<sup>2</sup>. This gives the following compact appearance observation model:

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

### 3.2.2. The Observation-to-Person Prior Distribution

The joint distribution of the assignment variables factorizes as:

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

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

$$p(Z_{tk}^i = n | \mathbf{e}_t) = e_{tn} a_{tn}^i \quad \text{with} \quad \sum_{n=0}^N e_{tn} a_{tn}^i = 1. \quad (8)$$

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 considering two hypothesis. Firstly the kinematic state dynamics follow a first-order Markov chain, meaning that the state  $\mathbf{X}_t$  only depends

---

<sup>2</sup>It should be noted that the normalization constant  $W_\lambda = \int_{\sum_k \mathbf{h}_k=1} \exp(-\lambda d_B(\mathbf{h}, \mathbf{h}_n)) d\mathbf{h}$  can be exactly computed only for histograms with dimension lower than 3. In practice  $W_\lambda$  is approximated using Monte Carlo integration.

293 on state  $\mathbf{X}_{t-1}$ . Secondly, the person locations do not influence each other's  
 294 dynamics, meaning that there is one first-order Markov chain for each person.  
 295 Formally, this can be written as:

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

296 The immediate consequence is that the posterior distribution computes:

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

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

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

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:

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

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

#### 304 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  $\Theta^\circ$ , the optimal approximation has the following general expression:

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

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

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:

$$\log q(Z_{tk}^i) = \mathbf{E}_{q(\mathbf{x}_t)} \{ \log p(Z_{tk}^i, \mathbf{X}_t | \mathbf{o}_{1:t}, \mathbf{e}_{1:t}, \Theta^\circ) \}, \quad (14)$$

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

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

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

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

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

#### 321 4.1. E-Z-Step

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

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

325 where

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

326 and  $\epsilon_{tkn}^i$  is defined as:

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

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

#### 332 4.2. E-X-Step

333 The estimation of  $q(\mathbf{X}_{tn})$  is derived from (13). Similarly to the previ-  
 334 ous posterior distribution, the mathematical derivations are provided in Ap-  
 335 pendix 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

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

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

336 We note that the variational approximation of the kinematic-state distribu-  
 337 tion reminds the Kalman filter solution of a linear dynamical system with  
 338 mainly one difference: in our solution (21) and (22), the mean vectors and  
 339 covariance matrices are computed with the observations weighted by  $\alpha_{tkn}^i$   
 340 (see (21) and (22)).

### 341 4.3. M-step

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

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

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

347 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  $\Sigma^i$ . This is done by maximizing

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

348 with respect to  $\Sigma^i$ . Differentiating  $J(\Sigma^i)$  with respect to  $\Sigma^i$  and equating to  
 349 zero gives:

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

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

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

350 Equating differential of the cost  $J(\Lambda_n)$  to zeros gives:

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

351 It is worth noticing that, in the current filtering formalism, the formulas  
 352 for  $\Sigma^i$  and  $\Lambda_n$  are instantaneous, i.e., they are estimated only from the

353 observations at time  $t$ . The information at time  $t$  is usually insufficient to  
 354 obtain stable values for these matrices. Even if estimating  $\Sigma^i$  and  $\Lambda_n$  is  
 355 suitable in a parameter learning scenario where the tracks are provided, we  
 356 noticed that in practical tracking scenarios, where the tracks are unknown,  
 357 this does not yield stable results. Suitable priors on the temporal dynamics of  
 358 the covariance parameters are required. Therefore, in this paper we assume  
 359 that the observation and dynamical model covariance matrices are fixed.

## 360 5. Person-Birth and Person-Visibility Processes

361 Tracking a time-varying number of targets requires procedures to cre-  
 362 ate tracks when new targets enter the scene and to delete tracks when  
 363 corresponding targets leave the visual scene. In this paper, we propose a  
 364 statistical-test based birth process that creates new tracks and a hidden  
 365 Markov model (HMM) based visibility process that handles disappearing tar-  
 366 gets. Until here, we assumed that the existence variables  $e_{tn}$  were given. In  
 367 this section we present the inference model for the existence variable based  
 368 on the stochastic birth-process.

### 369 5.1. Birth Process

370 The principle of the person birth process is to search for consistent tra-  
 371 jectories in the history of observations associated to clutter. Intuitively, two  
 372 hypotheses “*the considered observation sequence is generated by a person*  
 373 *not being tracked*” and “*the considered observation sequence is generated by*  
 374 *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  $\{\mathbf{y}_{t-L,k_L}, \dots, \mathbf{y}_{t,k_0}\}$ ,<sup>3</sup> then the observation likelihood is  $p(\mathbf{y}_{t-l,k_l} | \mathbf{x}_{t-l,n}) = g(\mathbf{y}_{t-l,k_l}; \mathbf{P}\mathbf{x}_{t-l,n}, \Sigma)$  and the person trajectory is governed by the dynamical model  $p(\mathbf{x}_{t,n} | \mathbf{x}_{t-1,n}) = g(\mathbf{x}_{t,n}; \mathbf{D}\mathbf{x}_{t-1,n}, \Lambda_n)$ . 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  $p_b(\mathbf{x}_{t-L,n}) = g(\mathbf{x}_{t-L,n}; \mathbf{m}_b, \Gamma_b)$ , which is approximatively equivalent

---

<sup>3</sup>In practice we considered  $L = 2$ , however, derivations are valid for arbitrary values of  $L$ .

to a uniform distribution over the image. Consequently, the joint observation distribution writes:

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

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

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

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

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

## 387 5.2. Person-Visibility Process

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

394 We denote the  $n$ -th person visibility (binary) variable by  $V_{tn}$ , meaning  
395 that the person is visible at time  $t$  if  $V_{tn} = 1$  and 0 otherwise. We assume the  
396 existence of a transition model for the hidden visibility variable  $V_{tn}$ . More  
397 precisely, the visibility state temporal evolution is governed by the transition

398 matrix,  $p(V_{tn} = j|V_{t-1,n} = i) = \pi_v^{\delta_{ij}}(1 - \pi_v)^{1-\delta_{ij}}$ , where  $\pi_v$  is the probability  
 399 to remain in the same state. To enforce temporal smoothness, the probability  
 400 to remain in the same state is taken higher than the probability to switch to  
 401 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  $\nu_{tn}$ . In other words, we need to estimate the filtering distribution  $p(V_{tn}|\nu_{1:tn}, e_{1:tn})$  which can be written as:

$$p(V_{tn} = v_{tn}|\nu_{1:t}, e_{1:tn}) = \frac{p(\nu_{tn}|v_{tn}, e_{tn}) \sum_{v_{t-1,n}} p(v_{tn}|v_{t-1,n}) p(v_{t-1,n}|\nu_{1:t-1,n}, e_{1:t-1})}{p(\nu_{tn}|\nu_{1:t-1,n}, e_{1:t})}, \quad (29)$$

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

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

405 Intuitively, when  $\nu_{tn}$  is high, the likelihood is large if  $v_{tn} = 1$  (person is  
 406 visible). The opposite behavior is found when  $\nu_{tn}$  is small. Importantly,  
 407 at each frame, because the visibility state is a binary variable, its filtering  
 408 distribution can be straightforwardly computed.

## 409 6. Experiments

### 410 6.1. Evaluation Protocol

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

417 The second dataset is constituted of four sequences classically used in  
 418 computer vision to evaluate multi-person tracking methods [18, 19]. Two se-  
 419 quences were selected from the MOT Challenge Dataset [39]:<sup>4</sup> TUD-Stadmitte

---

<sup>4</sup><http://motchallenge.net/>



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.

420 (9 persons, 179 frames) and PETS09-S2L1 (18 persons, 795 frames). The  
 421 third sequence is the TownCentre sequence (231 persons, 4500 frames) recorded  
 422 by the Oxford Active Vision Lab. The last one is ParkingLot (14 persons,  
 423 749) recorded by the Center for Research in Computer Vision of University  
 424 of Central Florida. TUD-Stadmitte records closely viewed full body pedestrians.  
 425 PETS09-S2L1 and ParkingLot features a dozen of far-viewed full  
 426 body pedestrians. TownCentre captures a very large number of far viewed  
 427 pedestrians. This evaluation dataset is diverse and large (more than 6000  
 428 frames) enough to give a reliable assessment of the multi-person tracking  
 429 performance measures. Figure 3 shows typical views of all the sequences.

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

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

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

452 The computational cost of the proposed model is mainly due the the ob-  
 453 servation extraction, namely the person detection. This process is known in  
 454 computer vision to be computationally intensive. However, there are pedes-  
 455 trian detectors that achieve real time performances [45]. The VEM part of  
 456 the tracking model, which involves only inversion of 6 by 6 matrices, is com-  
 457 putationally efficient and can be made real time. It converges in less than 10  
 458 steps.

## 459 6.2. Validation on the Cocktail Party Dataset

460 In the cocktail party dataset our model exploits upper body detections  
 461 obtained using [25] and face detections obtained using [26]. Therefore, we  
 462 have two types of observations, upper body **U** and face **F**. The hidden state  
 463 corresponds to the position and velocity of the upper body. The observa-  
 464 tion operator  $\mathbf{P}^U$  (see section 3.2.1) for the upper body observations simply  
 465 removes the velocity components of the hidden state. The observation oper-  
 466 ator  $\mathbf{P}^F$  for the face observations combines a projection removing the veloc-  
 467 ity components and an affine mapping (scaling and translation) transform-  
 468 ing face localization bounding boxes into the the upper body localization  
 469 bounding boxes. The appearance observations are concatenations of joint  
 470 hue-saturation color histograms of the torso split into three different regions,  
 471 plus the head region as shown in Fig.4(a).

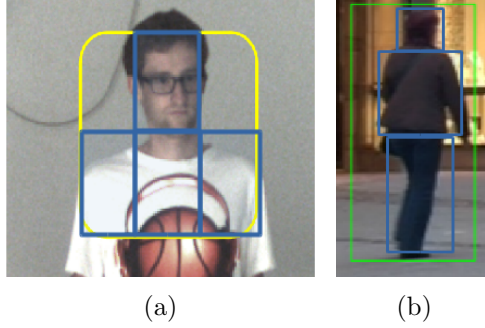


Figure 4: Region splitting for computing the color histograms: Fig.4(a) shows an example with upper-body detection while Fig.4(b) shows an example of full body detection.

Sequence	Features	<b>Rc</b>	<b>Pr</b>	<b>MOTA</b>	<b>MOTP</b>
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 %.

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

482 Results in Table 1 show that for the sequence CPD-2, while **Pr** and  
483 **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, 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.

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 cocktail party dataset.

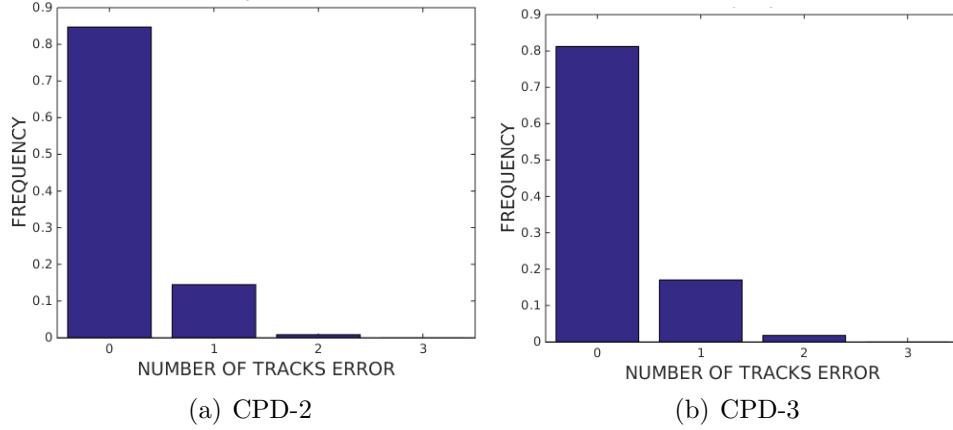


Figure 5: Histogram of absolute errors about the estimation of the number of people present in the visual scene over the Cocktail Party Dataset.

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

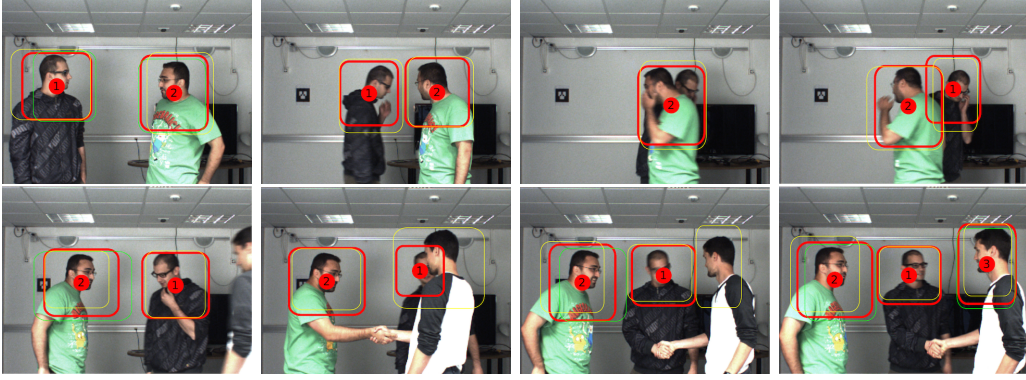


Figure 6: Sample tracking results on CPD-3. The green bounding boxes represent the face detections and the yellow bounding boxes represent the upper body detections. Importantly, the red bounding boxes display the tracking results.

the scene. More than 80% of the time, the correct 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). Ta-

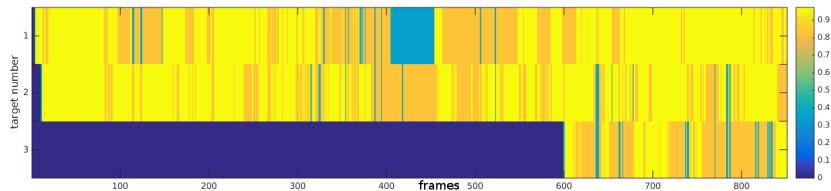


Figure 7: Estimated visibility probabilities for tracked persons in sequence CPD-3. Every row displays the corresponding targets visibility probabilities for every time frame. Yellow color represents very high probability (close to 1), and blue color represents very low probabilities.

Sequence	Method-Features	<b>Hausdorff</b>	<b>OMAT</b>	<b>OSPA</b>
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.

ble 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

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.

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 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.

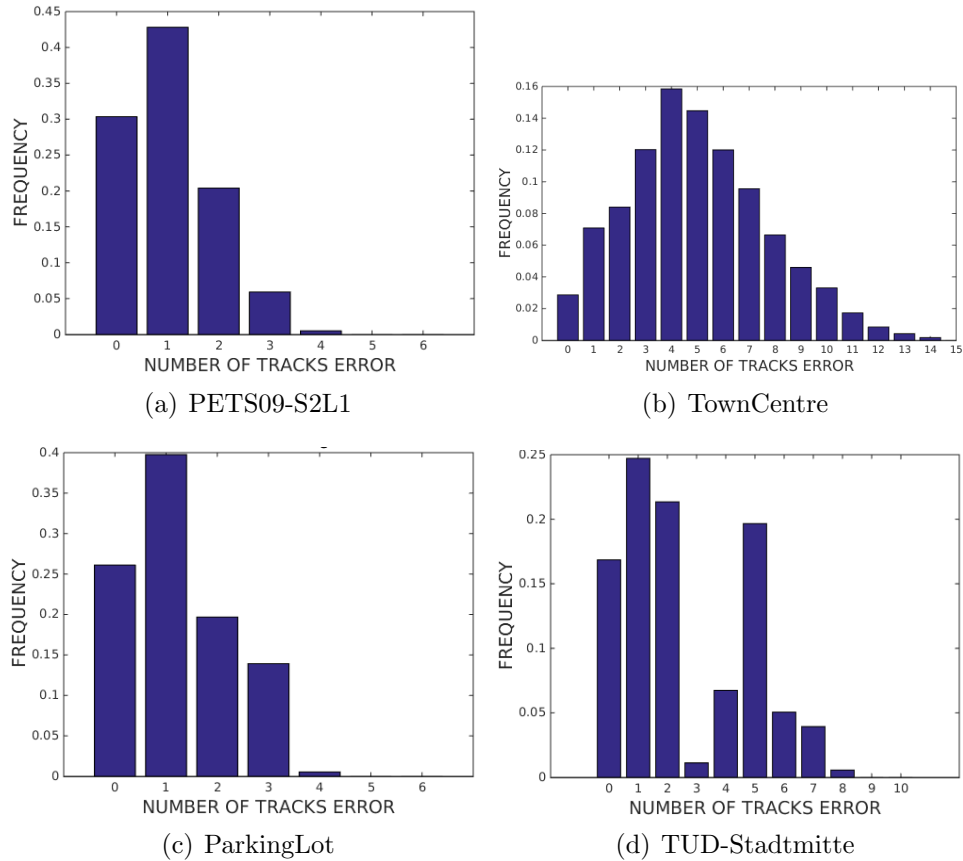


Figure 8: Histogram of errors about the estimation of the number of people present in the visual scene over ParkingLot, TownCentre, PETS09-S2L1, TUD-Stadtmitte.

583 For the TownCentre sequence which involves 231 persons over 4500 frames,  
 584 over 70% of the time, the error made by the variational tracker is below 7  
 585 persons. This shows that, even in challenging situations involving occlusions  
 586 due to crowd, the birth and the visibility process play their role.

587 Figure 9 presents sample results for the PET09-S2L1 sequence. In ad-  
 588 dition, videos presenting the results on the second dataset are provided as  
 589 supplementary material. These results show temporally consistent tracks.  
 590 Occasionally, person identity switches may occur when two people cross. Re-  
 591 markably, because the proposed tracking model is allowed to reuse the iden-  
 592 tity of persons visible in the past, people re-entering the scene after having  
 593 left, will be recognized the the previously used track will be awakened.



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.

## 594 7. Conclusions

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

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

## 618 Appendix A. Derivation of the Variational Formulation

### 619 Appendix A.1. Filtering Distribution Approximation

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

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

624 It is composed of three terms, the likelihood  $p(\mathbf{o}_t | \mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t)$ , the predictive  
 625 distribution  $p(\mathbf{Z}_t, \mathbf{X}_t | \mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})$ , and the normalization factor  $p(\mathbf{o}_t | \mathbf{o}_{1:t-1}, \mathbf{e}_{1:t})$

626 which is independent of the hidden variables. The likelihood can be expanded  
 627 as:

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

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

The predictive distribution factorizes as

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

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

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

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 be approximated by

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

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

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

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

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

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

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

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

639 and after some algebraic derivations, the distribution of interest can be writ-  
640 ten as the following multinomial distribution

$$q(Z_{tk}^i = n | \mathbf{e}_t) = \alpha_{tkn}^i = \frac{e_{tn} \epsilon_{tkn}^i}{\sum_{m=0}^N e_{tm} \epsilon_{tkm}^i} \quad (\text{A.6})$$

641 *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

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

642 where, as above,  $C$  gathers constant terms. After some algebraic derivation  
643 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  
644 covariance of the Gaussian distribution are given by (21) and by (22).

## 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.
- [6] M. Yang, Y. Liu, L. Wen, Z. You, A probabilistic framework for multi-target tracking with mutual occlusions, IEEE Conference 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).

- 674 [12] H. Sidenbladh, Multi-target particle filtering for the probability hypoth-  
 675 esis density, in: IEEE International Conference on Information Fusion,  
 676 Tokyo, Japan, 2003, pp. 800–806.
- 677 [13] D. Clark, J. Bell, Convergence results for the particle PHD filter, IEEE  
 678 Transactions on Signal Processing 54 (7) (2006) 2652–2661.
- 679 [14] B.-N. Vo, S. Singh, A. Doucet, Random finite sets and sequential monte  
 680 carlo methods in multi-target tracking, in: IEEE International Radar  
 681 Conference, Huntsville, USA, 2003, pp. 486–491.
- 682 [15] W. K. Ma, B. N. Vo, S. S. Singh, Tracking an unknown time-varying  
 683 number of speakers using TDOA measurements: a random finite set  
 684 approach, IEEE Transactions on Signal Processing 54 (9) (2006) 3291–  
 685 3304.
- 686 [16] E. Maggio, M. Taj, A. Cavallaro, Efficient multitarget visual tracking  
 687 using random finite sets, IEEE Transactions on Circuits and Systems  
 688 for Video Technology 18 (8) (2008) 1016–1027.
- 689 [17] B. Yang, R. Nevatia, An online learned CRF model for multi-target  
 690 tracking, in: IEEE Conference on Computer Vision and Pattern Recog-  
 691 nition, Providence, USA, 2012, pp. 2034–2041.
- 692 [18] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for  
 693 multitarget tracking, IEEE Transactions on Pattern Analysis and Ma-  
 694 chine Intelligence 36 (1) (2014) 58–72.
- 695 [19] A. Heili, A. Lopez-Mendez, J.-M. Odobez, Exploiting long-term con-  
 696 nectivity and visual motion in CRF-based multi-person tracking, IEEE  
 697 Transactions on Image Processing 23 (7) (2014) 3040–3056.
- 698 [20] Y. Bar-Shalom, F. Daum, J. Huang, The probabilistic data association  
 699 filter: estimation in the presence of measurement origin and uncertainty,  
 700 IEEE Control System Magazine 29 (6) (2009) 82–100.
- 701 [21] J. Vermaak, N. Lawrence, P. Perez, Variational inference for visual track-  
 702 ing, in: IEEE Conference on Computer Vision and Pattern Recognition,  
 703 Madison, USA, 2003, pp. 773–780.

- 704 [22] V. Smidl, A. Quinn, *The Variational Bayes Method in Signal Processing*,  
705 Springer, 2006.
- 706 [23] C. M. Bishop, *Pattern Recognition and Machine Learning (Information  
707 Science and Statistics)*, Springer, 2007.
- 708 [24] A. Yilmaz, O. Javed, M. Shah, Object tracking: A survey, in: *ACM  
709 Computing Surveys*, 2006, p. 13.
- 710 [25] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detec-  
711 tion with discriminatively trained part based models, *IEEE Transactions  
712 on Pattern Analysis and Machine Intelligence* 32 (9) (2010) 1627–1645.
- 713 [26] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark  
714 localization in the wild, in: *IEEE Computer Vision and Pattern Recog-  
715 nition*, Providence, USA, 2012, pp. 2879–2886.
- 716 [27] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature  
717 space analysis, *IEEE Transactions on Pattern Analysis and Machine  
718 Intelligence* 24 (5) (2002) 603–619.
- 719 [28] S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on par-  
720 ticle filters for online nonlinear/non-Gaussian bayesian tracking, *IEEE  
721 Transactions on Signal Processing* 50 (2) (2002) 174–188.
- 722 [29] S. Sarka, A. Vehtari, J. Lampinen, Rao-Blackwellized Monte Carlo data  
723 association for multiple target tracking, in: *IEEE International Confer-  
724 ence on Information Fusion*, Stockholm, Sweden, 2004, pp. 583–590.
- 725 [30] Y. Yan, A. Kostin, W. Christmas, J. Kittler, A novel data associa-  
726 tion algorithm for object tracking in clutter with application to tennis  
727 video analysis, in: *IEEE Computer Vision and Pattern Recognition*,  
728 New York, USA, 2006, pp. 634–641.
- 729 [31] S.-W. Bae, K.-J. Yoon, Robust online multi-object tracking based on  
730 tracklet confidence and online discriminative appearance learning, in:  
731 *IEEE Computer Vision and Pattern Recognition*, Columbus, USA, 2014,  
732 pp. 1218–1225.
- 733 [32] H. Pirsiavash, D. Ramanan, C. C. Fowlkes, Globally-optimal greedy  
734 algorithms for tracking a variable number of objects, *IEEE Conference  
735 on Computer Vision and Pattern Recognition* (2011) 1201–1208.

- 736 [33] M. Isard, J. MacCormick, Bramble: A bayesian multiple-blob tracker, in:  
737 IEEE International Conference on Computer Vision, British Columbia,  
738 Canada, 2001, pp. 34–41.
- 739 [34] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, D. G. Lowe, A  
740 boosted particle filter: Multitarget detection and tracking, in: European  
741 Conference on Computer Vision, Prague, Czech Republic, 2004, pp. 28–  
742 39.
- 743 [35] S. Sarka, A. Vehtari, J. Lampinen, Rao-Blackwellized particle filter for  
744 multiple target tracking, in: IEEE International Conference Information  
745 Fusion, Québec ,Canada, 2007, pp. 2–15.
- 746 [36] A. R. Zamir, A. Dehghan, M. Shah, Gmcp-tracker: Global multi-object  
747 tracking using generalized minimum clique graphs, IEEE Conference on  
748 Computer Vision and Pattern Recognition (2012) 343–356.
- 749 [37] W. Longyin, W. Li, J. Yan, Z. Lei, D. Yi, S. Z. Li, Multiple target  
750 tracking based on undirected hierarchical relation hypergraph, IEEE  
751 Conference on Computer Vision and Pattern Recognition (2014) 1282–  
752 1289.
- 753 [38] P. Perez, C. Hue, J. Vermaak, M. Gangnet, Color-based probabilistic  
754 tracking, in: European Conference on Computer Vision, Copenhagen,  
755 Denmark, 2002, pp. 661–675.
- 756 [39] L. Leal-Taixe, A. Milan, I. Reid, S. Roth, K. Schindler, MOT Challenge  
757 2015: Towards a benchmark for multi-target tracking, arXiv:1504.01942  
758 (2015).
- 759 [40] B. Ristic, B.-N. Vo, D. Clark, Performance evaluation of multi-target  
760 tracking using the OSPA metric, in: IEEE International Conference on  
761 Information Fusion, Edinburgh, UK, 2010, pp. 1–7.
- 762 [41] K. Smith, D. Gatica-Perez, J.-M. Odobez, S. Ba, Evaluating multi-  
763 object tracking, in: IEEE CVPR Workshop on Empirical Evaluation  
764 Methods in Computer Vision, San Diego, USA, 2005, pp. 36–36.
- 765 [42] R. Stiefelhagen, K. Bernardin, R. Bowers, J. S. Garofolo, D. Mostefa,  
766 P. Soundararajan, CLEAR 2006 evaluation, in: First International

- 767 Workshop on Classification of Events and Relationship, CLEAR 2006,  
768 Springer, 2005.
- 769 [43] W. Longyin, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H.  
770 Yang, S. Lyu, DETRAC filter multiple target tracker: A new benchmark  
771 and protocol for multi-object tracking, arXiv:1511.04136.
- 772 [44] D. E. Clark, K. Panta, B. N. Vo, The GM-PHD filter multiple target  
773 tracker, IEEE International Conference In Information Fusion (2006)  
774 1–8.
- 775 [45] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, D. Ferguson, Real-  
776 time pedestrian detection with deep network cascades, British Machine  
777 Vision Conference.
- 778 [46] R. Streit, Birth and death in multitarget tracking filters, in: IEEE Work-  
779 shop on Sensor Data Fusion: Trends, Solutions, Applications, 2013, pp.  
780 1–6.