Tracking Multiple Persons Based on a Variational Bayesian Model

Yutong Ban[1], Sileye Ba[2(✉)], Xavier Alameda-Pineda[3], and Radu Horaud[1]

[1] Inria Grenoble Rhône-Alpes, Montbonnot-saint-martin, France
[2] VideoStitch, Paris, France
sileye.ba@video-stitch.com
[3] University of Trento, Trento, Italy

Abstract. Object tracking is an ubiquitous problem in computer vision with many applications in human-machine and human-robot interaction, augmented reality, driving assistance, surveillance, etc. Although thoroughly investigated, tracking multiple persons remains a challenging and an open problem. In this paper, an online variational Bayesian model for multiple-person tracking is proposed. This yields a variational expectation-maximization (VEM) algorithm. The computational efficiency of the proposed method is due to closed-form expressions for both the posterior distributions of the latent variables and for the estimation of the model parameters. A stochastic process that handles person *birth* and person *death* enables the tracker to handle a varying number of persons over long periods of time. The proposed method is benchmarked using the MOT 2016 dataset.

1 Introduction

The problem of object tracking is ubiquitous in computer vision. While many object tracking methods are available, multiple-person tracking remains extremely challenging [1]. In addition to the difficulties related 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, for the following reasons: (i) the observations associated with detectors need to be associated to objects being tracked, which includes the process of discarding detection errors, (ii) the number of objects is not known in advance and hence it must be estimated and updated over time, (iii) mutual occlusions (not present in single-tracking scenarios) must be robustly handled, and (iv) 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, an exact recursive filtering solution is intractable.

Support from the ERC Advanced Grant VHIA number 340113 and from MIUR Active Aging at Home CTN01 00128 is greatly acknowledged.

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part II, LNCS 9914, pp. 52–67, 2016.
DOI: 10.1007/978-3-319-48881-3_5

Several multiple-person tracking methods have been proposed within the trans-dimensional Markov chain model [2], where the dimensionality of the state-space is treated as a state variable. This allows to track a variable number of objects by jointly estimating the number of objects and their states. [3–5] 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 [6]. The random finite set framework proposed in [7–9] is also very popular, where 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 computationally intensive, an approximation known as the probability hypothesis density (PHD) filter was proposed [10]. Further sampling-based approximations of random-set based filters were subsequently proposed, e.g. [11–13]. These were exploited in [14] for tracking a time-varying number of active speakers using auditory cues and in [15] for multiple-target tracking using visual observations. Recently, conditional random fields have been introduced to address multiple-target tracking [16–18]. In this case, tracking is cast into an energy minimization problem. In radar tracking, popular multiple-target tracking methods are joint probabilistic data association (JPDA), and multiple hypothesis filters [19].

An interesting and less investigated framework for multiple-target tracking is the variational Bayesian class of models for tracking an unknown and varying number of persons. Although variational models are very popular in machine learning, their use for object tracking has been limited to tracking a fixed number of targets [20]. Variational Bayes methods approximate the joint a posteriori distribution of the complete set of latent variables by a separable distribution [21,22]. In an online tracking scenario, where only past and current observations are available, this leads to approximating the filtering distribution. An interesting aspect of variational methods is that they yield closed-form expressions for the posterior distributions of the hidden variables and for the model parameters, thus enabling an intrinsically efficient filtering procedure implemented via a variational EM (VEM) algorithm. In this paper, we derive a variational Bayesian formulation for multiple-person tracking, and present results on the MOT 2016 challenge dataset [23]. The proposed method extends [24] in many apsects: (i) the assignment variables are included in the filtering equation and therefore the state variables and the assignment variables are jointly inferred, (ii) a temporal window is incorporated in the visibility process, leading to a tracker that is more robust to misdetections, (iii) death process allows to forget about *old* tracks and thus opens the door to large-scale processing, as needed in many realistic situations. Finally, full evaluation of the proposed tracker within the MOT 2016 challenge dataset assesses its performance against other state-of-the-art methods in a principled and systematic way. Examples of results obtained with our method and Matlab code are publicly available.[1]

The remainder of this paper is organized as follows. Section 2 details the proposed Bayesian model and a variational solution is presented in Sect. 3. In Sect. 4,

[1] https://team.inria.fr/perception/research/ovbt/.

we depict the birth, visibility and death processes allowing to handle an unknown and varying number of persons. Section 5 presents benchmarking results. Finally, Sect. 6 draws conclusions.

2 Variational Multiple-Person Tracking

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.

Let N be the maximum number of persons. A track $n \leq N$ 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 Sects. 3 and 4; Their inference, grounded in a birth stochastic process, is discussed in Sect. 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 and size, 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 multiple-person state random vector is denoted by $\mathbf{X}_t = (\mathbf{X}_{t1}^\top, \ldots, \mathbf{X}_{tN}^\top)^\top \in \mathbb{R}^{6N}$.

Fig. 1. Examples of detected persons from the MOT 2016 dataset.

We assume the existence of a person detector, providing K_t localization observations at each time t. The k-th localization observation delivered by the detector at time t is denoted by $\mathbf{y}_{tk} \in \mathbb{R}^4$, and represents the location (2D position, width, height) of a person, e.g. Figure 1. The set of observations at time t is denoted by $\mathbf{y}_t = \{\mathbf{y}_{tk}\}_{k=1}^{K_t}$. Associated to $\mathbf{y}_{tk}$, there is a photometric description of the person appearance, denoted by $\mathbf{h}_{tk}$. This photometric observation is extracted from the bounding box of $\mathbf{y}_{tk}$. Altogether, the localization and photometric observations constitute the observations $\mathbf{o}_{tk} = (\mathbf{y}_{tk}, \mathbf{h}_{tk})$ used by our tracker. Definitions analogous to $\mathbf{y}_t$ hold for $\mathbf{h}_t = \{\mathbf{h}_{tk}\}_{k=1}^{K_t}$ and $\mathbf{o}_t = \{\mathbf{o}_{tk}\}_{k=1}^{K_t}$. The probability of a set of random variables is written as $p(\mathbf{o}_t) = p(\mathbf{o}_{t1}, \ldots, \mathbf{o}_{tK_t})$.

We also define an observation-to-person assignment (hidden) variable Z_{tk}, associated with each observation $\mathbf{o}_{tk}$. $Z_{tk} = n, n \in \{1 \dots N\}$ means that $\mathbf{o}_{tk}$ is associated to person n. 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} is extended to $\{0\} \cup \{1 \dots N\}$, and $Z_{tk} = 0$ means that observation $\mathbf{o}_{tk}$ 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 an estimation robust to outliers.

2.1 The Online Tracking Model

The online multiple-person tracking problem is cast into the estimation of the filtering distribution of the hidden variables given the causal observations $p(\mathbf{Z}_t, \mathbf{Z}_{t-1}, \mathbf{X}_t, \mathbf{X}_{t-1} | \mathbf{o}_{1:t}, \mathbf{e}_{1:t})$, where $\mathbf{o}_{1:t} = \{\mathbf{o}_1, \dots, \mathbf{o}_t\}$. Importantly, we assume that the observations at time t only depend on the hidden and visibility variables at time t. The filtering distribution can be written as:

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

The denominator of (1) 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 two terms of the latter, namely the observation model $p(\mathbf{o}_t | \mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t)$ and the dynamic distribution $p(\mathbf{Z}_t, \mathbf{X}_t | \mathbf{Z}_{t-1}, \mathbf{X}_{t-1}, \mathbf{e}_t)$.

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

$$p(\mathbf{o}_t | \mathbf{Z}_t, \mathbf{X}_t, \mathbf{e}_t) = \prod_{k=1}^{K_t} p(\mathbf{o}_{tk} | Z_{tk}, \mathbf{X}_t, \mathbf{e}_t). \quad (2)$$

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:

$$\begin{aligned} p(\mathbf{o}_{tk} | Z_{tk}, \mathbf{X}_t, \mathbf{e}_t) &= p(\mathbf{y}_{tk}, \mathbf{h}_{tk} | Z_{tk}, \mathbf{X}_t, \mathbf{e}_t) \\ &= p(\mathbf{y}_{tk} | Z_{tk}, \mathbf{X}_t, \mathbf{e}_t) p(\mathbf{h}_{tk} | Z_{tk}, \mathbf{e}_t). \end{aligned} \quad (3)$$

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

The linear operator $\mathbf{P}$ maps the kinematic state vectors onto the space of observations. For example, when $\mathbf{X}_{tn}$ represents the full-body kinematic state (full-body localization and velocity) and $\mathbf{y}_{tk}$ represents the full-body localization observation, $\mathbf{P}$ is a projection which, when applied to a state vector, only retains the localization components of the state vector. Finally, the full observation model is compactly defined by the following, where δ_{ij} stands for the Kronecker function:

$$p(\mathbf{y}_{tk}|Z_{tk} = n, \mathbf{X}_t, \mathbf{e}_t) = u(\mathbf{y}_{tk})^{1-e_{tn}} \left(u(\mathbf{y}_{tk})^{\delta_{0n}} g(\mathbf{y}_{tk};\ \mathbf{P}\mathbf{X}_{tn}, \mathbf{\Sigma})^{1-\delta_{0n}}\right)^{e_{tn}}. \quad (4)$$

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

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

where λ is a positive skewness parameter, $d_B(\cdot)$ is the Battacharya distance between histograms, $\mathbf{h}_n$ is the reference appearance model of person n. This gives the following compact appearance observation model:

$$p(\mathbf{h}_{tk}|Z_{tk} = n, \mathbf{X}_t, \mathbf{e}_t) = u(\mathbf{h}_{tk})^{1-e_{tn}} (u(\mathbf{h}_{tk})^{\delta_{0n}} b(\mathbf{h}_{tk}; \mathbf{h}_n)^{1-\delta_{0n}})^{e_{tn}}. \quad (5)$$

The Dynamic Distribution. Here we consider two hypotheses, firstly, we assume the at each time instance, assignment variable doesn't depends on the previous assignment. So we can factorize the the dynamic distribution into the observation-to-person prior distribution and the predictive distribution. Secondly, the kinematic state dynamics follow a first-order Markov chain, meaning that the state $\mathbf{X}_t$ only depends on state $\mathbf{X}_{t-1}$.

$$p(\mathbf{Z}_t, \mathbf{X}_t|\mathbf{Z}_{t-1}, \mathbf{X}_{t-1}, \mathbf{e}_t) = p(\mathbf{Z}_t|\mathbf{e}_t) p(\mathbf{X}_t|\mathbf{X}_{t-1}, \mathbf{e}_t). \quad (6)$$

The Observation-to-Person Prior Distribution. The joint distribution of the assignment variables can be factorized as:

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

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

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

Because e_{tn} takes the value 1 only for actual persons, the probability to assign an observation to a non-existing person is null. When person n is visible, a_{tn} represents the probability of observation $\mathbf{y}_{tk}$ to be generated from person n.

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 $p(\mathbf{X}_t|\mathbf{X}_{t-1}, \mathbf{e}_t)$. The predictive distribution is mainly driven by the dynamics of persons's kinematic states, which are modeled assuming that 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:

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

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:

$$p(\mathbf{X}_{tn} = \mathbf{x}_{tn}|\mathbf{X}_{t-1,n} = \mathbf{x}_{t-1,n}, e_{tn}) = u(\mathbf{x}_{tn})^{1-e_{tn}} g(\mathbf{x}_{tn};\ \mathbf{D}\mathbf{x}_{t-1,n}, \mathbf{\Lambda}_n)^{e_{tn}}, \tag{10}$$

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:

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

In other words, the dynamics of an existing person n, *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 $\mathbf{\Theta} = \big(\{\mathbf{\Sigma}\},\ \{\mathbf{\Lambda}_n\}_{n=1}^{N}, \mathbf{A}_{1:t}\big)$, with $\mathbf{A}_t = \{a_{tn}\}_{n=0}^{N}$.

3 Variational Bayesian Inference

Because of the combinatorial nature of the observation-to-person assignment problem, a direct optimization of the filtering distribution (1) 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 $p(\mathbf{Z}_t, \mathbf{Z}_{t-1}, \mathbf{X}_t, \mathbf{X}_{t-1}|\mathbf{o}_{1:t}, \mathbf{e}_{1:t})$ by a separable distribution, e.g. $q(\mathbf{Z}_t)\prod_{n=0}^{N} q(\mathbf{X}_{tn})$. According to the variational Bayesian formulation [21,22], given the observations and the parameters at the previous iteration $\mathbf{\Theta}^{\circ}$, the optimal approximation has the following general expression:

$$\log q(\mathbf{Z}_t) = \mathbf{E}_{q(\mathbf{X}_t)q(\mathbf{X}_{t-1})q(\mathbf{Z}_{t-1})}\left\{\log \widetilde{P}\right\}, \tag{11}$$

$$\log q(\mathbf{Z}_{t-1}) = \mathbf{E}_{q(\mathbf{X}_t)q(\mathbf{X}_{t-1})q(\mathbf{Z}_t)}\left\{\log \widetilde{P}\right\}, \tag{12}$$

$$\log q(\mathbf{X}_{tn}) = \mathbf{E}_{q(\mathbf{Z}_t)q(\mathbf{Z}_{t-1})q(\mathbf{X}_{t-1,n})\prod_{m\neq n} q(\mathbf{X}_{tm})}\left\{\log \widetilde{P}\right\}, \tag{13}$$

$$\log q(\mathbf{X}_{t-1,n}) = \mathbf{E}_{q(\mathbf{Z}_t)q(\mathbf{Z}_{t-1})q(\mathbf{X}_{t,n})\prod_{m\neq n} q(\mathbf{X}_{t-1,m})}\left\{\log \widetilde{P}\right\}, \tag{14}$$

where, for simplicity, we used the notation $\widetilde{P} = p(\mathbf{Z}_t, \mathbf{Z}_{t-1}, \mathbf{X}_t, \mathbf{X}_{t-1}|\mathbf{o}_{1:t}, \mathbf{e}_{1:t}, \mathbf{\Theta}^\circ)$. In our particular case, when these two equations are put together with the probabilistic model defined in (2), (6) and (9), the expression of $q(\mathbf{Z}_t)$ is factorized further into:

$$\log q(Z_{tk}) = \mathbf{E}_{q(\mathbf{X}_t)q(\mathbf{X}_{t-1})q(\mathbf{Z}_{t-1})}\left\{\log \widetilde{P}\right\}, \tag{15}$$

Note that this equation leads to a finer factorization that the one we initially 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. (2), (6) and (9). The final factorization writes:

$$p(\mathbf{Z}_t, \mathbf{Z}_{t-1}, \mathbf{X}_t, \mathbf{X}_{t-1}|\mathbf{o}_{1:t}, \mathbf{e}_{1:t}) \approx \prod_{k=0}^{K_t} q(Z_{tk}) \prod_{k=0}^{K_{t-1}} q(Z_{t-1,k}) \prod_{n=0}^{N} q(\mathbf{X}_{tn}) q(\mathbf{X}_{t-1,n}). \tag{16}$$

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

$$J(\mathbf{\Theta}, \mathbf{\Theta}^\circ) = \mathbf{E}_{q(\mathbf{Z},\mathbf{X})}\left\{\log p(\mathbf{Z}_t, \mathbf{Z}_{t-1}, \mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{o}_{1:t}|\mathbf{e}_{1:t}, \mathbf{\Theta}, \mathbf{\Theta}^\circ)\right\}. \tag{17}$$

3.1 E-Z-Step

The estimation of $q(Z_{tk})$ is carried out by developing the expectation (15) which yields the following formula:

$$q(Z_{tk} = n) = \alpha_{tkn} = \frac{e_{tn}\epsilon_{tkn}a_{tn}}{\sum_{m=0}^{N} e_{tm}\epsilon_{tkm}a_{tn}}, \tag{18}$$

and ϵ_{tkn} is defined as:

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

where $\mathrm{Tr}(\cdot)$ is the trace operator and $\boldsymbol{\mu}_{tn}$ and $\mathbf{\Gamma}_{tn}$ 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.

3.2 E-X-Step

The estimation of $q(\mathbf{X}_{tn})$ is derived from (13). Similarly to the previous posterior distribution, which boil down to the following formula:

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

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

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

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

Similarly, for the estimation of the distribution

$$q(\mathbf{X}_{t-1,n}) = u(\mathbf{X}_{t-1,n})^{1-e_{tn}} g(\mathbf{X}_{t-1,n}; \widehat{\boldsymbol{\mu}}_{t-1,n}, \widehat{\boldsymbol{\Gamma}}_{t-1,n})^{e_{tn}}, \tag{23}$$

the mean and covariance are:

$$\widehat{\boldsymbol{\Gamma}}_{t-1,n} = \left(\mathbf{D}^\top \boldsymbol{\Lambda}_n^{-1} \mathbf{D} + \boldsymbol{\Gamma}_{t-1,n} \right)^{-1} \tag{24}$$

$$\widehat{\boldsymbol{\mu}}_{t-1,n} = \widehat{\boldsymbol{\Gamma}}_{t-1,n} \left(\mathbf{D}^\top \boldsymbol{\Lambda}_n^{-1} \boldsymbol{\mu}_{t,n} + \boldsymbol{\Gamma}_{t-1,n}^{-1} \boldsymbol{\mu}_{t-1,n} \right). \tag{25}$$

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 formulation, (21) and (22), the means and covariances are computed by weighting the observations with α_{tkn}, i.e. (21) and (22).

3.3 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 (17). Concerning the parameters of the a priori observation-to-object assignment $\mathbf{A}_t$ we compute:

$$J(a_{tn}) = \sum_{k=1}^{K_t} e_{tn} \alpha_{tkn} \log(e_{tn} a_{tn}) \quad \text{s.t.} \quad \sum_{n=0}^{N} e_{tn} a_{tn} = 1, \tag{26}$$

and we trivially obtain:

$$a_{tn} = \frac{e_{tn} \sum_{k=1}^{K_t} \alpha_{tkn}}{\sum_{m=0}^{N} e_{tm} \sum_{k=1}^{K_t} \alpha_{tkm}}. \tag{27}$$

The observation covariance $\boldsymbol{\Sigma}$ and the state covariances $\boldsymbol{\Lambda}_n$ can be estimated during the M-step. However, in our current implementation estimates for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}_n$ are instantaneous, i.e., they are obtained only from the observations at time t (see the experimental section for details).

4 Person-Birth, -Visibility and -Death 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 model for the existence variable based on the stochastic birth-process.

4.1 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 are confronted, namely: *(i) the considered observation sequence is generated by a person not being tracked* and *(ii) the considered observation sequence is generated by clutter*.

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 (4) and (10). 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:

$$\begin{aligned}\tau_0 &= p(\mathbf{y}_{t,k_0}, \ldots, \mathbf{y}_{t-L,k_L}) \\ &= \int p(\mathbf{y}_{t,k_0}, \ldots, \mathbf{y}_{t-L,k_L}, \mathbf{x}_{t:t-L,n}) d\mathbf{x}_{t:t-L,n} \\ &= \int \prod_{l=0}^{L} p(\mathbf{y}_{t,k_l}|\mathbf{x}_{t-l,n}) \times \prod_{l=0}^{L-1} p(\mathbf{x}_{t-l,n}|\mathbf{x}_{t-l-1,n}) \times p_b(\mathbf{x}_{t-2,n}) d\mathbf{x}_{t:t-L,n}, \quad (28)\end{aligned}$$

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 (4). 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

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

$$\tau_1 = p(\mathbf{y}_{t,k_0}, \ldots, \mathbf{y}_{t-L,k_L}) = \prod_{l=0}^{L} u(\mathbf{y}_{t-l,k_l}). \tag{29}$$

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 $e_{tn} = 1, q(\mathbf{x}_{t,n}; \boldsymbol{\mu}_{t,n}, \boldsymbol{\Gamma}_{t,n})$ with $\boldsymbol{\mu}_{t,n} = [\mathbf{y}_{t,k_0}^\top, \mathbf{0}_2^\top]^\top$, and $\boldsymbol{\Gamma}_{tn}$ is set to the value of a birth covariance matrix (see (20)). Also, the reference appearance model for the new person is defined as $\mathbf{h}_{t,n} = \mathbf{h}_{t,k_0}$.

4.2 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 whenever their reference appearance 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 $\nu_{tn} = e_{tn}a_{tn}$, 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:

$$p(V_{tn} = v_{tn}|\nu_{1:t}, e_{1:tn}) = \frac{p(\nu_{tn}|v_{tn}, e_{tn}) \sum_{v_{t-1,n}} p(v_{tn}|v_{t-1,n}) p(v_{t-1,n}|\nu_{1:t-1,n}, e_{1:t-1})}{p(\nu_{tn}|\nu_{1:t-1,n}, e_{1:t})}, \tag{30}$$

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

$$p(\nu_{tn}|v_{tn}, e_{tn}) = (\exp(-\lambda\nu_{tn}))^{v_{tn}}(1 - \exp(-\lambda\nu_{tn}))^{1-v_{tn}} \tag{31}$$

Intuitively, when ν_{tn} is high, the likelihood is large if $v_{tn} = 1$ (person is visible). The opposite behavior is found when ν_{tn} is small. Importantly, at each frame, because the visibility state is a binary variable, its filtering distribution can be straightforwardly computed. We found this rather intuitive strategy to be somewhat "shaky" over time even taking the Markov dependency into account. This is why we enriched the visibility observations to span over multiple frames $\nu_{tn} = \sum_{l=0}^{L} e_{t+ln}a_{t+ln}$, so that if $v_{tn} = 1$, the likelihood is large when ν_{tn} is high and therefore the target is visible in one or more neighboring frames. This is the equivalent of the hypothesis testing spanning over time associated to the birth process.

4.3 Death Process

The idea of the person-visibility process arises from encouraging track consistency when a target disappears and appears back in the field of view. However, a tracker that remembers *all* the tracks that have been previously seen is hardly scalable. Indeed, the memory resources required by a system that remembers all previous appearance templates grows indefinitely with new appearances. Therefore, one must discard *old* information to facilitate the scalability of the method to large datasets containing sequences with several dozens of different people involved. In addition to alleviating the memory requirements, this also reduces the computational complexity of the tracker. This is the motivation of including a death process into the proposed variational framework. Intuitively one would like to discard those tracks that have not been seen during several frames. In practice, we found that discarding those tracks that are not visible for ten consecutive frames yields a good trade-off between complexity, resource demand and performance. Setting this parameter for a different dataset should not be chimeric, since the precise interpretation of the meaning of it is straightforward.

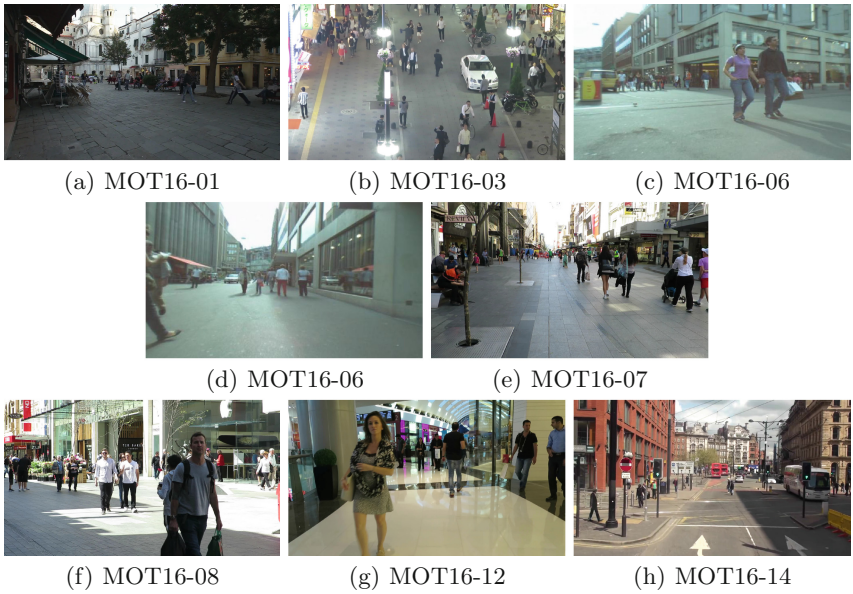
5 Experiments, Performance Evaluation, and Benchmark

We evaluated the performance of the proposed variational multiple-person tracker on the MOT 2016 dataset challenge [23]. This dataset is composed of seven training videos and seven test videos. Importantly, we use the detections that are provided with the dataset. Because multiple-person tracking intrinsically implies track creation (birth), deletion (death), target identity maintenance, and localization, evaluating multiple-person tracking techniques is a non-trivial task. Many metrics have been proposed, e.g. [25–28].

We adopt the metrics used by the MOT 2016 benchmark, namely [27]. The main tracking measures are: the *multiple-object tracking accuracy* (MOTA), that combines false positives (FP), missed targets (FN), and identity switches (ID); the *multiple-object tracking precision* (MOTP), that measures the alignment of the tracker output bounding box with the ground truth; the false alarm per frame (FAF); the ratio of mostly tracked trajectories (MT); the ratio of mostly lost trajectories (ML) and the number of track fragmentations (Frag).

Figure 2 shows sample images of all test videos: They contain three sequences recorded with static cameras (MOT16-01, MOT16-03 and MOT16-08), which contain very crowded scenes and thus are very challenging, and five sequences with large camera motions, both translations and rotations, which make the data even more difficult to process.

As explained above, we use the public pedestrian detections provided within the MOT16 challenge. These static detections are complemented in two different ways. First, we extract velocity observations by means of a simple optical-flow based algorithm that looks for the most similar region of the next temporal frame within the neighborhood of the original detection. Therefore, the observations operator P is the identity matrix, project the entire state variable into the



(a) MOT16-01 (b) MOT16-03 (c) MOT16-06

(d) MOT16-06 (e) MOT16-07

(f) MOT16-08 (g) MOT16-12 (h) MOT16-14

Fig. 2. Samples images from the MOT 16 test sequences.

observation space. Second, the appearance feature vector is the concatenation of joint color histograms of three regions of the torso in HSV space.

The proposed variational model is governed by several parameters. Aiming at providing an algorithm that is dataset-independent and that features a good trade-off between flexibility and performance, we set the observation covariance matrix $\mathbf{\Sigma}$ and the state covariance matrix $\mathbf{\Lambda}_n$ automatically from the detections. More precisely, both matrices are imposed to be diagonal; for $\mathbf{\Sigma}$, the variances of the horizontal position, of the width, and of the horizontal speed are $1/3$, $1/3$ and $1/6$ of the detected width. The variances of the vertical quantities are built analogously. The rationale behind this choice is that we consider that the true detection lies, more or less, within the width and height of the detected bounding box. Regarding $\mathbf{\Lambda}_n$, the diagonal entries are $1, 1$ and $1/2$ of the detected width, and vertical quantities are defined analogously. Furthermore, in order to eliminate arbitrary false detections, we set $L = 5$ in the birth process. Finally, for sequences in which the size of the bounding boxes is roughly constant, we discarded those detections that were too large or too small.

Examples of the tracking results for all the test sequences except MOT16-07 are shown in Fig. 3, while six frames from MOT16-07 are shown in Fig. 4. In all figures, the red boxes represent our tracking result and the numbers within the boxes are the tracking indexes. Generally speaking, on one hand the variational model is crucial to properly associate detections with trajectories. On the other hand, the birth and visibility processes play a role when tracked objects appear and disappear. Regarding Fig. 4, it contains 54 tracks recorded by a moving

(a) MOT16-01 (b) MOT16-03 (c) MOT16-06

(d) MOT16-08 (e) MOT16-12 (f) MOT16-14

Fig. 3. Sample results on several sequences of MOT16 datasets, red bounding boxes represents the tracking results, and the number inside each box is the track index. (Color figure online)

camera in a sequence of 500 frames. It is a very challenging tracking task, not only because the density of pedestrians is quite high, but also because significant camera motion makes the person trajectories to be both rough and discontinuous. One drawback of the proposed approach is that partially consistent false detections could lead to the creation of a false track, therefore tracking an inexistent pedestrian. On the positive side, the main advantage of the proposed model is that the probabilistic combination of the dynamic and appearance models can decrease the probability of switching the identities of two tracks.

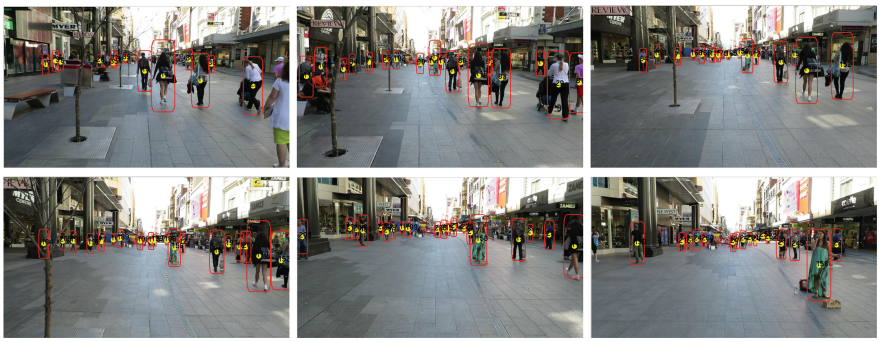


Fig. 4. Sample results on the sequence MOT16-07, encoded as in the previous figure. (Color figure online)

Table 1 reports the performance of the proposed algorithm, which is referred to as OVBT (online variational Bayesian tracker), over the seven test sequences

Table 1. Evaluation of the proposed multiple-person tracking method with different features on the seven sequences of the MOT16 test dataset.

Sequence	MOTA	MOTP	FAF	MT	ML	FP	FN	ID Sw	Frag
MOT16-01	23.9	71.4	1.5	13.0 %	39.1 %	696	4,137	35	89
MOT16-03	46.9	75.7	4.1	17.6 %	20.3 %	6,173	48,631	689	1,184
MOT16-06	32.7	73.2	0.5	3.6 %	58.4 %	562	7,073	124	183
MOT16-07	33.6	73.3	2.2	9.3 %	35.2 %	1,077	9,605	158	272
MOT16-08	24.6	78.4	1.7	3.2 %	41.3 %	1,066	11,402	150	177
MOT16-12	32.8	76.7	0.9	10.5 %	52.3 %	766	4,749	63	80
MOT16-14	18.1	74.5	1.6	2.4 %	61.6 %	1,177	13,866	102	155
Over All	38.4± 8.8	75.4	1.9	7.5 %	47.3 %	11,517	99,463	1,321	2,140

of the MOT 2016 challenge. The results obtained with OVBT are available on the MOT 2016 webpage.[3] One can notice that our method provides high precision (MOTP) but low accuracy (MOTA), meaning that some tracks were missed (mostly due to misetections). This is consistent with a rather low MT measure. This behavior was more extreme when the visibility process did not include any observation aggregation over time. Indeed, we observed that considering multiple observations within the visibility process leads to better performance (for all sequences and almost all measures).

6 Conclusions

We propose a variational Bayesian solution to the multiple-target problem. In the literature, other solutions based on sampling such as MCMC, and random finite set, such as the PHD filter have been proposed to solve the same problem. Comparison with other state of the art methods are available [24].

The main goal of our study was to benchmark the model on MOT Challenge 2016. Implementation issues of the tracker are discussed as well as its strengths and weaknesses regarding the absolute performance on the test sequences and the relative performance when compared with other participants to the MOT Challenges.

The presented model is free from magic parameters, since these are automatically derived from the data. Moreover, the proposed temporal aggregation for the visibility process appears to be an excellent complement to the variational Bayes EM algorithm. In the near future, we plan to derive self-paced learning strategies within this variational framework able to automatically assess which detections must be used for tracking and which should not be utilized.

[3] https://motchallenge.net/results/MOT16/.

References

1. Luo, W., Xing, J., Zhang, X., Zhao, W., Kim, T.K.: Multiple object tracking: a review (2015). arXiv:1409.761
2. Green, P.J.: Trans-dimensional Markov chain Monte Carlo. In: Oxford Statistical Science Series, pp. 179–198 (2003)
3. Khan, Z., Balch, T., Dellaert, F.: An MCMC-based particle filter for tracking multiple interacting targets. In: European Conference on Computer Vision, Prague, Czech Republic, pp. 279–290 (2004)
4. Smith, K., Gatica-Perez, D., Odobez, J.M.: Using particles to track varying numbers of interacting people. In: IEEE Computer Vision and Pattern Recognition, San Diego, USA, pp. 962–969 (2005)
5. Yang, M., Liu, Y., Wen, L., You, Z.: A probabilistic framework for multitarget tracking with mutual occlusions. In: IEEE Confenrence on Computer Vision and Pattern Recognition, pp. 1298–1305 (2014)
6. Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**(4), 711–732 (1995)
7. Mahler, R.P.: Multisource multitarget filtering: a unified approach. In: Aerospace/Defense Sensing and Controls, International Society for Optics and Photonics, pp. 296–307 (1998)
8. Mahler, R.P.S.: Statistics 101 for multisensor, multitarget data fusion. IEEE Aerosp. Electron. Syst. Mag. **19**(1), 53–64 (2004)
9. Mahler, R.P.S.: Statistics 102 for multisensor multitarget data fusion. IEEE Sel. Top. Sign. Proces. **19**(1), 53–64 (2013)
10. Mahler, R.P.S.: A theoretical foundation for the Stein-Winter "probability hypothesis density (PHD)" multitarget tracking approach. Technical report (2000)
11. Sidenbladh, H.: Multi-target particle filtering for the probability hypothesis density. In: IEEE International Conference on Information Fusion, Tokyo, Japan, pp. 800–806 (2003)
12. Clark, D., Bell, J.: Convergence results for the particle PHD filter. IEEE Trans. Sign. Proces. **54**(7), 2652–2661 (2006)
13. Vo, B.N., Singh, S., Doucet, A.: Random finite sets and sequential monte carlo methods in multi-target tracking. In: IEEE International Radar Conference, Huntsville, USA, pp. 486–491 (2003)
14. Ma, W.K., Vo, B.N., Singh, S.S.: Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach. IEEE Trans. Sign. Proces. **54**(9), 3291–3304 (2006)
15. Maggio, E., Taj, M., Cavallaro, A.: Efficient multitarget visual tracking using random finite sets. IEEE Trans. Circ. Syst. Video Technol. **18**(8), 1016–1027 (2008)
16. Yang, B., Nevatia, R.: An online learned CRF model for multi-target tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, pp. 2034–2041 (2012)
17. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. IEEE Trans. Pattern Anal. Mach. Intell. **36**(1), 58–72 (2014)
18. Heili, A., Lopez-Mendez, A., Odobez, J.M.: Exploiting long-term connectivity and visual motion in CRF-based multi-person tracking. IEEE Trans. Image Process. **23**(7), 3040–3056 (2014)
19. Bar-Shalom, Y., Daum, F., Huang, J.: The probabilistic data association filter: estimation in the presence of measurement origin and uncertainty. IEEE Control Syst. Mag. **29**(6), 82–100 (2009)

20. Vermaak, J., Lawrence, N., Perez, P.: Variational inference for visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, Madison, USA, pp. 773–780 (2003)
21. Smidl, V., Quinn, A.: The Variational Bayes Method in Signal Processing. Springer, Heidelberg (2006)
22. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, New York (2007)
23. Milan, A., Leal-Taix, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. In: arXiv:1603.00831 [cs] (2016)
24. Ba, S., Alameda-Pineda, X., Xompero, A., Horaud, R.: An on-line variational Bayesian model for multi-person tracking from cluttered scenes. Computer Vision and Image Understanding (2016)
25. Ristic, B., Vo, B.N., Clark, D.: Performance evaluation of multi-target tracking using the OSPA metric. In: IEEE International Conference on Information Fusion, Edinburgh, UK, pp. 1–7 (2010)
26. Smith, K., Gatica-Perez, D., Odobez, J.M., Ba, S.: Evaluating multi-object tracking. In: IEEE CVPR Workshop on Empirical Evaluation Methods in Computer Vision, San Diego, USA, pp. 36–36 (2005)
27. Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., Soundararajan, P.: The CLEAR 2006 evaluation. In: Stiefelhagen, R., Garofolo, J. (eds.) CLEAR 2006. LNCS, vol. 4122, pp. 1–44. Springer, Heidelberg (2007). doi:10.1007/978-3-540-69568-4_1
28. Longyin, W., Du, D., Cai, Z., Lei, Z., Chang, M.C., Qi, H., Lim, J., Yang, M.H., Lyu, S.: DETRAC filter multiple target tracker: a new benchmark and protocol for multi-object tracking, arXiv:1511.04136 (2015)