Variational Inference and Learning of Piecewise-linear Dynamical Systems
Xavier Alameda-Pineda, Vincent Drouard, Radu Horaud

To cite this version:
Xavier Alameda-Pineda, Vincent Drouard, Radu Horaud. Variational Inference and Learning of Piecewise-linear Dynamical Systems. IEEE Transactions on Neural Networks and Learning Systems, IEEE, 2022, 33 (8), pp.3753 - 3764. 10.1109/TNNLS.2021.3054407. hal-02745527v3

HAL Id: hal-02745527
https://hal.archives-ouvertes.fr/hal-02745527v3
Submitted on 26 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract—Modeling the temporal behavior of data is of primordial importance in many scientific and engineering fields. Baseline methods assume that both the dynamic and observation equations follow linear-Gaussian models. However, there are many real-world processes that cannot be characterized by a single linear behavior. Alternatively, it is possible to consider a piecewise-linear model which, combined with a switching mechanism, is well suited when several modes of behavior are needed. Nevertheless, switching dynamical systems are intractable because their computational complexity increases exponentially with time. In this paper, we propose a variational approximation of piecewise linear dynamical systems. We provide full details of the derivation of two variational expectation-maximization algorithms, a filter and a smoother. We show that the model parameters can be split into two sets, static and dynamic parameters, and that the former parameters can be estimated off-line together with the number of linear modes, or the number of states of the switching variable. We apply the proposed method to the head-pose tracking, and we thoroughly compare our algorithms with several state of the art trackers.

Index Terms—Switching state space model, linear dynamical system, inverse regression, Bayesian inference, variational approximation, expectation-maximization, Kalman filter.

I. INTRODUCTION

Modeling the temporal behavior of data is of primordial importance in many scientific fields, such as signal processing [1], computer vision [2], [3], robotics [4], autonomous navigation [5], to cite just a few. The baseline model for addressing this problem is the linear dynamical system (LDS). The basic idea of LDS is to assume that both the dynamic and the observation equations follow linear-Gaussian models. This yields tractable learning and inference procedures, namely the Kalman filter (KF) [6] and the Kalman smoother (KS) [7].

In many situations, the latent (state) space, whose dynamics must be modeled, is embedded in a high-dimensional observed space. In general, the direct mapping, from the low-dimensional latent space to the high-dimensional feature (observed) space, as well as the inverse of this mapping, are non-linear. Moreover, the dynamics of the latent space may be non-linear as well. Several methods were proposed to deal with non-linear dynamical systems, e.g. Bayesian tracking with particle filters [8], [9], the extended Kalman filter (EKF) [10], and the unscented Kalman filter (UKF) [11]. Alternatively, it is possible to consider several linear dynamic models and to combine them with a switching mechanism that selects over time one among several linear regimes: this is referred to as the switching Kalman filter (SKF), the switching LDS, the jump-Markov linear system, or the switching state space model [12].

The mixture Kalman filter (MKF) [8] uses a sequential Monte Carlo method based on a random mixture of Gaussians to approximate the target distribution. It formulates non-linear systems into conditional or partial conditional LDSs. Outcomes of non-linear/non-Gaussian Bayesian trackers based on sequential importance sampling are reviewed and discussed in [9], most notably the problems of degeneracy, choice of importance density, and resampling. The basic idea of EKF is to linearize the equations using a first-order Taylor expansion and to apply the standard KF to the linearized model. The additional error due to linearization is not taken into account which may lead to sub-optimal performance. Rather than approximating a non-linear dynamical system with a linear one, UKF specifies the state distribution using a minimal set of deterministically selected sample points. The sample points, when propagated through the true non-linear system, capture the posterior state distribution accurately to the third order Taylor expansion. The stability of UKF was thoroughly investigated in the control literature, e.g. [13]. It was shown that the design of the noise covariance, of both state and observation equations, critically affects the performance of the filter.

The methods just outlined generally deal with a single non-linear or linearized, state equation. There are many real-world processes that cannot be characterized with a single state equation, but with multiple discrete modes of behavior, both in terms of their dynamics and of their observation model, in particular when the latter must predict (generate) high-dimensional observations from a low-dimensional state space.

We consider the problem of tracking the orientation of a person head/face (three rotation angles) from a sequence of images, referred to as head-pose tracking (HPT). A face detector provides input to a face descriptor immune to illumination changes, background conditions, as well as inter- and intra-person variabilities (shape and aspect). Face descriptors of choice are histograms of oriented gradients (HOGs) [14] and embeddings based on convolutional neural networks (CNNs) [15]–[18]. These high-dimensional feature vectors contain head pose information implicitly and a number of non-linear or piecewise-linear regression methods have been proposed to extract head pose, namely Gaussian process regression [19], support vector regression [20], kernel partial least squares [21],
deep inverse regression [18], or Gaussian mixture of linear regressions [22], [23].

Recently it was proposed to approximate non-linear high-dimensional to low-dimensional (high-to-low) mappings with mixtures of linear-Gaussian [22], [24] and linear-Student [25] regressions. These models adopt an inverse regression strategy, namely they learn a low-to-high mapping followed by the evaluation of a high-to-low mapping. The rationale of this way of doing is manifold: (i) low-to-high regression learning avoids the estimation of a large number of parameters, hence it requires a relatively small amount of training data, (ii) the parameters of the high-to-low regression are analytically evaluated from the low-to-high parameters, (iii) the mixture model setting has the advantage of providing inference procedures using closed-form EM algorithms. It is interesting to note that these Gaussian/Student mixtures group data with similar regression associations together. Within the same cluster, the association can be considered as locally linear, which can then be resolved under the classical linear regression setting. This piecewise linear models are well suited to capture potentially non-linear relations. This was extensively discussed in [22] and in [24], and was successfully applied to both head-pose estimation [23] and audio-source localization [26], [27].

In this paper we propose a variational expectation-maximization algorithm to learn piecewise-linear dynamic systems (P-LDSs). A P-LDS can be viewed either as a piecewise-linear approximation of a non-linear dynamic system [9], or as a dynamic generalization of mixtures of linear regressions [22], [23], [26]. A P-LDS may also be viewed as soft version of switching LDS [12], [28]. The assignment variable of the piecewise-linear mixture model plays the role of the switching variable of both the dynamic and observation models and it is governed by the transition matrix of a corresponding hidden Markov model (HMM). It is well known that the complexity of these hybrid dynamical systems, i.e. systems that combine discrete- and continuous-valued latent variables, increases exponentially with time [12]. Indeed, for $K$ linear models and after $T$ time steps, the exact marginalized posterior distribution of the state is a Gaussian mixture with $K^2$ components. Therefore, the problem of learning the parameters of such hybrid systems must be carried out via approximate solutions. Traditionally, inference of hybrid dynamical systems is based on approximating the posterior distribution with a simpler one, e.g. the generalized pseudo Bayes filter. In this paper we propose an alternative based on replacing the difficult-to-compute posterior with an approximate tractable posterior.

The remainder of this paper is organized as follows. Section II describes related work. Section III formulates P-LDS and analyses its intractability. Section IV describes in detail the proposed variational approximation model as well as the two EM algorithms, a variational filter and a variational smoother. For the sake of completeness, Section V describes a GPB2 approximation of P-LDS. Section VI describes experimental results obtained with head-pose tracking.  

I. RELATED WORK

The intractability of switching LDS (and of P-LDS) can be addressed using sampling methods: sequential Monte Carlo methods (particle filtering) have been used for this purpose. The main drawback is that sampling can be inefficient, which leads to slow convergence. To reduce the size of the state space, Rao-Blackwellisation may be employed. Instead of drawing the samples from the joint posterior of the discrete and continuous states, tractable sub-structures of the model can be utilized. Non parametric Bayesian inference of switching LDS was proposed in [29], where a switching mechanism is used for the dynamic model, while the observation model uses a standard LDS. This is problematic when the observation model is not linear. Moreover, a Gibbs sampler is used for inference, providing asymptotic properties tied to a high computational cost.

A general theory of Rao-Blackwellised particle filters (RBPFs) applied to dynamic Bayesian networks (DBNs) was proposed in [30]. In the case of switching LDS, marginalization of the joint posterior, namely analytic integration over the continuous variables, considerably reduces the sampling space. RBPF using Gibbs sampling was used for speech recognition [31], where the discrete state corresponds to phonemes and the continuous state corresponds to a time evolving representation of the observations. RBPF using Metropolis-Hastings sampling was used in [32] to analyse motion patterns of bees.

In [32] it was noticed that a naive exploration of the space of discrete variables is prohibitive and that data-driven (DD) MCMC sampling improves convergence. DD-MCMC requires supervised learning from a labeled training dataset, which may be cumbersome, if not prohibitive, because it is not practical to manually associate discrete-variable values with the observed vectors. More recently, [33] proposed a mixture of switching LDSs to analyse the dynamic behavior of pedestrians: an MCMC inference scheme uses both Gibbs and Metropolis-Hastings samplers. While potentially powerful, MCMC methods and their variants are non-analytic methods and typically suffer from slow convergence rates (they are only exact in the case of infinite size samples), especially in high-dimensional spaces, which is impractical in the case of tracking. A recent comparison between MCMC and variational inference emphasizes that the latter easily takes advantage of stochastic and distributed optimization [34].

The generalized pseudo Bayes (GPB) [35] and the GPB of order two (GPB2) [36], [37] algorithms belong to the assumed density filtering (ADF) [38] class of models which is widely used to approximate an intractable distribution with a simpler one. GPB collapses the mixture of $K$ Gaussian components, resulting from considering all the possible linear models at each time step, into one Gaussian component. GPB2 is more sophisticated and more time consuming than GPB, as it collapses the $K^2$ Gaussian components, resulting from considering all the possible linear models when going from one time step to the next, into $K$ Gaussians components. The GPB2 algorithm was applied to the analysis of motor cortical activity of hand movements in macaque monkeys.
[1], and more recently it was used for tracking eye gaze [39] and for path prediction of pedestrians in the context of intelligent vehicles [5]. An offline extension of GBP2, called expectation correction was proposed in [40] and applied to speech recognition robust to noise [41].

Alternatively, structural variational inference and learning techniques consider a parameterized distribution which is in some sense close to the desired posterior distribution, as well as easier to compute. Variational models modify the structure of the posterior by removing dependencies between variables, i.e. the joint posterior distribution $P$ is approximated by a tractable variational distribution $Q$ with variational parameters $\theta$. The Kullback-Leibler divergence between $Q$ and $P$ is minimized with respect to $\theta$. In the case of switching LDSs, [42], [28] and [43] propose to remove some of the dependencies between the continuous and discrete latent variables, thus yielding tractable solvers. The mixed-state DBN proposed in [42] is an HMM driving the LDS bias. In [28] an HMM switches between several LDSs, each LDS having a different latent variable with its own dynamic regime. Both [42] and [28] lead to an EM algorithm whose maximization step (learning) satisfies a set of fixed-point equations in the variational parameters. The variational model proposed in [43] is more general than the one proposed in [42] and in [28]. Nevertheless, their approximation breaks the dependencies between the HMM and the LDS as well as the temporal dependencies.

The variational model of [43] was applied to speech recognition [44] and to speech production [45], while the model of [42] was applied to human motion capture [46]. It is interesting to note that in spite of the recent popularity gained by variational models, e.g. [47], as they provide tractable solutions to various intractable inference problems, e.g. [2], [3], [48]– [56], at the best of our knowledge, variational inference of switching LDS has not been addressed for the last decade.

More recently, learning and inference of dynamical systems have been addressed in the framework of deep generative models (DGMs), where the linear-Gaussian transition and emission distributions of LDS are replaced with non-linear Gaussian models. In detail, the mean and covariance of a Gaussian distribution are modeled with neural networks. Because of this non-linear dependencies, direct optimization of the corresponding data log-likelihood function is intractable. This issue is solved by maximization of a variational lower bound of the log-likelihood. For example, [57] uses a recurrent neural network (RNN) to model the mean and diagonal covariance matrix. The proposed structured inference network corresponds to a deep Kalman smoother, that needs both past and future observations. This formulation belongs to a wider class of non-linear Gaussian state-space models that were recently reviewed in [58]. Deep neural networks can also be used within structured variational inference for pixel-level prediction tasks [59], [60], but we are not aware of any works addressing switching LDS with this methodology.

With respect to the related work just outlined, this paper has the following contributions. We propose a variational approximation of P-LDS. Unlike the variational model of [28], which switches between several linear regimes, we propose a piecewise-linear model that approximates non-linear dynamic models. Unlike [43], the proposed variational approximation doesn’t break the temporal dependencies. Unlike the RNN-based non-linear state-space model of [57], the proposed variational piecewise-linear model yields a closed-form solver and it has fewer parameters.

We provide full details of an EM algorithm that can be differently used either as a variational filter or as a variational smoother. Unlike the models of [42] and of [28] that lead to solving a set of fixed-point equations, we develop a closed-form solution for learning the model parameters. Moreover, the proposed method benefits from a closed-form EM algorithm for the off-line estimation of the static parameters, namely, those associated with the observation model. This has two practical outcomes: (i) it reduces the task of learning to the estimation of the parameters of the dynamic model and (ii) it allows to learn the number of states of the switching variable (or, equivalently, the number of linear models) based on a model selection principle, i.e. the Bayes information criterion, [22], [25]. For the sake of completeness, we describe a GBP2 algorithm for P-LDS that slightly differs from the standard GBP2 for switching LDS in that it only needs to estimate the parameters of the dynamic model.

### III. Problem Formulation

Let $X \in \mathcal{X} \subset \mathbb{R}^L$ be a latent (or state) random variable and $Y \in \mathcal{Y} \subset \mathbb{R}^D$ be an observation variable. Without loss of generality it will be assumed that the dimensionality of the observation space is much higher than the dimensionality of the latent space, $D \gg L$. Let $x$ and $y$ denote realizations of $X$ and $Y$, respectively. Let $t \in \mathbb{N}$ be the discrete time index: $X_t$ denotes the latent variable at $t$ and the notation $X_{1:t}$ is a shorthand for the temporal sequence $X_1, \ldots, X_t$. In an LDS, the observed vectors are connected to the latent vectors through an observation equation. We will consider the following piecewise-linear observation model. It is assumed that at any time $t$ a realization $(y_t, x_t)$ of $(Y_t, X_t)$ is such that $y_t$ is generated from $x_t$ by a linear function $y_t = l_k(x_t) + \text{error term}$.

At each time step $t$, a discrete latent variable $Z_t$ is introduced, such that $Z_t = k$ if and only if $y_t$ is the image of $x_t$ by $l_k$, with $k \in \{1, \ldots, K\} \subset \mathbb{N}$. The piecewise linear function that maps the state $X_t$ onto the observation $Y_t$ is:

$$y_t = \sum_{k=1}^{K} \mathbb{I}(Z_t = k)(A_{Z_t}x_t + b_{Z_t} + e_{Z_t}),$$

where matrix $A_{Z_t=k} \in \mathbb{R}^{D \times L}$ and vector $b_{Z_t=k} \in \mathbb{R}^{D}$ define $l_k$, $\mathbb{I}(\cdot)$ is the indicator function and $e \sim \mathcal{N}(0, \Sigma)$ is a zero-mean Gaussian noise vector with covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$. The description of the model is completed by a similar piecewise linear dynamic equation:

$$x_t = \sum_{k=1}^{K} \mathbb{I}(Z_t = k)(C_{Z_t}x_{t-1} + w_{Z_t}),$$

The matrix $C_{Z_t=k} \in \mathbb{R}^{L \times L}$ and the vector $w_{Z_t=k} \in \mathbb{R}^{L}$ define $l_k$.
where \( C_{Z_t} \in \mathbb{R}^{L \times L} \) is the state transition matrix and \( w \sim \mathcal{N}(0, Q) \) is a zero-mean Gaussian noise vector with covariance matrix \( Q \in \mathbb{R}^{L \times L} \). To summarize, the \( k \)-th LDS is defined by the following probabilistic model, see Fig. 1:

\[
\begin{align*}
  p(x_t | x_{t-1}, Z_t = k) &= \mathcal{N}(x_t; C_{Z_t}x_{t-1}, Q_{Z_t}), \\
  p(y_t | x_t, Z_t = k) &= \mathcal{N}(y_t; A_{Z_t}x_t + b_{Z_t}, \Sigma_{Z_t}), \\
  p(x_1 | Z_t = k) &= \mathcal{N}(x_1; \gamma_{Z_t}, \Gamma_{Z_t}), \\
  p(Z_1 = k) &= \pi_{Z_1},
\end{align*}
\]

where \( \{\gamma_k, \Gamma_k, \pi_k\}_{k=1}^K \) is a zero-mean Gaussian mixture with a number of components that increases exponentially with time, i.e. there are \( K = 10 \) (the number of linear-Gaussian models). This distribution can be obtained by marginalization over the continuous and discrete variables given the observations:

\[
p(x_t | y_{1:t}) = \sum_{z_{1:t-1}} p(x_t, x_{1:t-1}, z_{1:t} | y_{1:t}) dx_{1:t-1},
\]

where \( \sum_{z_{1:t-1}} \int dx_{1:t-1} \) are shorthands for \( \int_{z_{1:t-1}} \sum_{z_{t-1} = 1}^K \int dx_{t-1} \) and \( dx_{1:t-1} \) for \( dx_1 \ldots dx_{t-1} \), respectively. Using the first-order Markov dependencies shown in the graphical model of Fig. 1, the joint probability (right hand side of (10)) can be factorized as:

\[
\begin{align*}
  p(x_t, z_{1:t} | y_{1:t}) &= \prod_{t'=2}^t p(y_{t'}, x_{t'}, z_{t'}) p(x_{t'} | x_{t'-1}, z_{t'}), \\
  &\times \int p(y_{1:t'} | x_{1:t'}) dx_{1:t'} p(x_{t'} | x_{t'-1}, z_{t'}), \\
\end{align*}
\]

Substituting the factors of (11) with their expressions (3)-(7) and using standard properties of Gaussian distributions and using marginalization, the joint distribution (11) writes:

\[
p(\hat{x}_t, x_{1:t-1}, z_{1:t}, y_{1:t}) = \beta_t \mathcal{N}(\hat{x}_t; x_t; \kappa_t, K_t),
\]

where the notation \( \hat{x}_t = [x_1 \ldots x_t] \) denotes vertical concatenation of vectors \( x_1, \ldots, x_t \), and where the weight \( \beta_t \), mean \( \kappa_t \), and covariance \( K_t \) depend on the model parameters (8) and (9) and on \( Z_{1:t} \). Therefore, the predictive distribution (10) is a Gaussian mixture with a number of components that increases exponentially with time, i.e. there are \( K_t \) components after \( t \) time steps, which is intractable.

This phenomenon appears not only when attempting to evaluate \( p(x_t | y_{1:t}) \) (filtering), but as well as when \( p(x_{1:T} | y_{1:T}) \) (smoothing) is evaluated. While the former is used for on-line via model selection, using the Bayes information criterion (BIC), or empirically, based on the mean absolute error between the predicted outputs and the ground-truth values, i.e. [23].

The parameters \( \phi \) characterize the dynamic behavior of both the continuous (3) and discrete (7) state variables and hence they must be estimated from a temporal sequence of observations. It is interesting to note that when the observation space is of high dimension, the strategy consisting of independently estimating the parameters \( \theta \) and \( \phi \) simplifies the tasks of prediction and filtering.

Fig. 1. Graphical representation of the switching linear dynamic model that shows the dependencies between the latent variables (white circles) and the observed variables (gray circles).
prediction, i.e., when the model is already trained, the latter is part of the E-step of any algorithm used for learning the parameters, and therefore equally essential. In the following, we present our variational approximation to perform inference as well as the complete VEM algorithm for learning. We also discuss the derivation of the GPB2 algorithm [36] in the context of the proposed formulation.

IV. VARIATIONAL INFERENCE AND LEARNING

In this section we present a variational approximation of P-LDS and we derive a variational EM algorithm with tractable inference (expectation) and closed-form parameter learning (maximization). We assume that the continuous and discrete variables are independent, a posteriori: consequently, the joint distribution over $x_{1:T}$ and $z_{1:T}$ is approximated with the following factorization:

$$p(x_{1:T}, z_{1:T} | y_{1:T}) \approx q(z_{1:T}) q(x_{1:T}).$$  (13)

This follows the same philosophy as the factorial hidden Markov models [61]. However, here we deal with hybrid states, namely discrete and continuous, therefore the derivation [61] does not apply and we need to derive a new EM algorithm. Notice that the proposed model is different than the model of [28]. Indeed, the latter switches between several continuous states, with their own linear dynamic regimes, while the proposed model approximates a possibly non-linear dynamic regime with a piecewise-linear model, similar to GPB2, i.e. (44).

As with HMM and LDS learning, we assume that the entire sequence of observations is available for training and the challenge consists of inferring the entire chain of state variables and of estimating the model parameters, namely the parameter vectors $\theta$, i.e. (8) (observation model) and $\phi$, i.e. (9) (dynamic model). As already outlined in Section III, the parameters $\theta$ don’t depend on time and they can be estimated using the algorithm of [22]. Therefore, we only need to estimate the dynamic parameters $\phi$. Formally, we maximize the expected complete-data log-likelihood:

$$Q(\phi) = \mathbb{E}_{p(x_{1:T}, z_{1:T} | y_{1:T})} \left[ \log p(x_{1:T}, z_{1:T} | y_{1:T}; \phi) \right],$$  (14)

where the posterior distribution $p(x_{1:T}, z_{1:T} | y_{1:T})$ is evaluated with the model parameters at the previous iteration $\phi^{old}$, implicit in the previous equation to simplify the reading.

A. Inference: The Expectation Steps

The two posterior distributions (13) write:

$$\log q(z_{1:T}) = \mathbb{E}_{q(x_{1:T})} \left[ \log p(x_{1:T}, z_{1:T} | y_{1:T}) \right] + \text{const},$$  (15)

$$\log q(x_{1:T}) = \mathbb{E}_{q(z_{1:T})} \left[ \log p(x_{1:T}, z_{1:T} | y_{1:T}) \right] + \text{const},$$  (16)

These distributions are alternatively evaluated, as explained in detail below.

1) E-Z step: By developing (15), ignoring the constant terms and using (3)-(7) we obtain:

$$q(z_{1:T}) \propto N(y_{1:T}; \eta_{1} + b_{z_{1:T}}, \Sigma_{z_{1:T}}) \times \exp \left( -\frac{1}{2} \text{tr} \left( A_{z_{1:T}}^{-1} \Sigma_{z_{1:T}}^{-1} A_{z_{1:T}}^\top V_{1} \right) \right) \times N(\eta_{1}; \gamma_{z_{1:T}}, \Gamma_{z_{1:T}}) \exp \left( -\frac{1}{2} \text{tr} \left( \Gamma_{z_{1:T}}^{-1} V_{1} \right) \right) \times \prod_{t \geq 2} \left[ N(y_{1:T}; A_{z_{1:T}} \eta_{t} + b_{z_{1:T}}, \Sigma_{z_{1:T}}) \times \exp \left( -\frac{1}{2} \text{tr} \left( A_{z_{1:T}}^{-1} \Sigma_{z_{1:T}}^{-1} A_{z_{1:T}}^\top V_{t} \right) \right) \times N(\eta_{t}; C_{z_{1:T}} \eta_{t-1}, Q_{z_{1:T}}) \times \exp \left( -\frac{1}{2} \text{tr} \left( C_{z_{1:T}}^{-1} C_{z_{1:T}}^\top V_{t-1} + Q_{z_{1:T}}^{-1} V_{t} \right) \right) \times \exp \left( \text{tr} \left( Q_{z_{1:T}}^{-1} C_{z_{1:T}} W_{t} \right) \right) \tau_{z_{t-1:T}} \right),$$  (17)

where $\eta_{t} = \mathbb{E}_{q(x_{1:T})}[x_{t}]$ is the posterior mean of $x_{t}$, $V_{t} = \mathbb{E}_{q(x_{1:T})}[(x_{t}, x_{t})] - \eta_{t} \eta_{t}^{\top}$ is the posterior covariance of $x_{t}$ and $W_{t} = \mathbb{E}_{q(x_{1:T})}[(x_{t-1}, x_{t})] - \eta_{t-1} \eta_{t}^{\top}$ is the posterior cross-covariance of $x_{t-1}$ and $x_{t}$.

One may notice that it is possible to group the terms that depend on $z$ in (17), thus yielding:

$$q(z_{1:T}) \propto \rho_{1:z_{1:T}}^{1:z_{1:T}} \prod_{t \geq 2} \left( \rho_{t, z_{t}, \tau_{z_{t-1:T}}} \right),$$  (18)

which is equivalent to an HMM with observation probabilities $\rho_{1:z_{1:T}}^{1:z_{1:T}}$. Therefore, by computing the standard forward-backward algorithm for HMMs, one can easily obtain the forward $\rho_{t, z_{t}}^{1:z_{1:T}}$ and backward $\rho_{t, z_{t}}^{1:z_{1:T}}$ probabilities to eventually obtain the posteriori probability $p(z_{t} | y_{1:T}) : q(z_{t})$:

$$q(z_{t}) \propto \rho_{t, z_{t}}^{1:z_{1:T}} \times \rho_{t, z_{t}}^{1:z_{1:T}} \sum_{z} \rho_{z_{t}, z_{t}}^{1:z_{1:T}}.$$  (19)

The estimation of the transition parameters $\tau_{z_{t-1:T}}$ requires the joint posteriori probability distribution of $z_{t-1}, z_{t}$ which can be easily computed as:

$$q(z_{t-1:T}) = \rho_{t-1, z_{t-1:T}} \times \rho_{t-1, z_{t-1:T}} \tau_{z_{t-1:T}} \rho_{t, z_{t}}^{1:z_{1:T}}.$$  (20)

2) E-X step: By developing (16), ignoring the constant terms and using (3)-(7) we obtain:

$$\log q(x_{1:T}) \propto$$

$$-\frac{1}{2} \left( x_{t}^\top V_{1}^{-1} x_{t} - 2 x_{t}^\top V_{1}^{-1} \eta_{t} + x_{t}^\top \Gamma_{1}^{-1} x_{t} - 2 x_{t}^\top \Gamma_{1}^{-1} \eta_{t} \right)$$

$$+ \sum_{t \geq 2} \left( x_{t-1}^\top V_{t-1}^{-1} x_{t-1} - 2 x_{t-1}^\top V_{t-1}^{-1} \eta_{t} + x_{t-1}^\top Q_{t}^{-1} x_{t-1} - 2 x_{t-1}^\top Q_{t}^{-1} \eta_{t} \right),$$  (21)
where the following definitions hold:

\[ \eta_t^i = V_t^i \left( \sum_z \rho_{t,z} A_z^T \Sigma_z^{-1} (y_t - b_z) \right), \]

(22)

\[ (V_t^i)^{-1} = \sum_z \rho_{t,z} A_z^T \Sigma_z^{-1} A_z, \]

(23)

\[ \overline{\gamma} = \overline{\Gamma} \sum_z \rho_{t,z} \Gamma_z^{-1} \gamma_z, \]

(24)

\[ (\overline{\Sigma}_t)^{-1} = \sum_z \rho_{t,z} \Sigma_z^{-1}, \]

(25)

\[ \overline{S}_t^{-1} = \sum_z \rho_{t,z} C_z^T \Sigma_z^{-1} C_z, \]

(26)

\[ \overline{B}_t = \sum_z \rho_{t,z} B_z C_z, \]

(27)

\[ (\overline{B}_t)^{-1} = \sum_z \rho_{t,z} B_z C_z^{-1} B_z. \]

(28)

To be valid, the last equation requires that \( \rho_{t,z} C_z^T Q_{z}^{-1} C_z \) is invertible for all values of \( z \), this implies that \( C_z \) is a full rank matrix for all values of \( z \), which is a very mild assumption. This is very close to a standard LDS (Kalman filter), but different enough in that standard forward-backward recursions cannot be applied. Indeed, in a standard LDS the following relationship holds: \( \overline{R}_t^i \overline{Q}_{t-1}^{-1} \overline{R}_t = \overline{S}_{t-1}^{-1} \), which is not the case in general. This condition is equivalent to impose the same Gaussian dynamic model to all the realizations of \( Z_t \) (which clearly corresponds to a Kalman filter). However, from (21), one may easily see that the joint posterior distribution is a high-dimensional Gaussian, and therefore the marginals will also be Gaussian. Even if the relationship does not correspond to a standard LDS, it is still possible to write the forward-backward equations that efficiently solve in an exact manner the inference of \( q(x_t) \).

With the above notations the forward and backward recursions write, respectively:

\[ \eta_t^i = V_t^i \left( (V_t^i)^{-1} \eta_{t+1}^i + \overline{R}_t \overline{S}_t (\overline{S}_t + (V_{t-1}^i)^{-1})^{-1} \eta_{t-1}^i \right), \]

(29)

\[ \eta_t^i = V_t^i \left( V_t^i)^{-1} + (\overline{Q}_t)^{-1} - \overline{R}_t (\overline{S}_t)^{-1} + (V_{t-1}^i)^{-1} \right)^{-1} \eta_{t-1}^i, \]

and:

\[ \eta_t^i = V_t^i \overline{R}_t \Sigma_t^{-1} ((V_{t+1}^i)^{-1} \eta_{t+1}^i + (V_{t+1}^i)^{-1} \eta_{t+1}^i), \]

\[ (V_{t-1}^i)^{-1} = (V_{t+1}^i)^{-1} + (\overline{Q}_t)^{-1} + (V_{t-1}^i)^{-1} \overline{R}_t \Sigma_t^{-1} \]

\[ \Omega_{t+1}^{-1} = (V_{t+1}^i)^{-1} + (\overline{Q}_t)^{-1} \]

\[ \Omega_t^{-1} = \Omega_{t+1}^{-1} \]

The forward is initialized with \( (V_1^i)^{-1} = \overline{\Gamma}^{-1} + (V_1^i)^{-1} \) and \( \eta_1^i = V_1^i \left( \overline{\Gamma}^{-1} \eta_1 + (V_1^i)^{-1} \right) \). The backward recursion starts at \( t = T - 1 \) with \( (V_T^i)^{-1} = 0 \) (and therefore the value of \( \eta_T^i \) has no effect). Together, they allow us to write the posterior probability of \( x_t \):

\[ q(x_t | y_{1:T}) = \mathcal{N}(x_t; \eta_t, V_t), \]

(30)

\[ \eta_t = V_t \left( (V_t)^{-1} \eta_t^i + (V_t)^{-1} \eta_t^i \right), \]

(31)

\[ (V_t)^{-1} = (V_t)^{-1} + (V_t)^{-1}. \]

In order to estimate the parameters of the dynamics, \( \mathbf{C}_z \) and \( \mathbf{Q}_z \), one needs the joint posterior probability of \( x_t \) and \( x_{t-1} \):

\[ q(x_t; x_{t-1}; \eta_t^i, V_t^i) = \mathcal{N}(x_t; \eta_t, V_t), \]

(32)

\[ \eta_t = V_t \left( (V_t)^{-1} \eta_t^i + (V_t)^{-1} \eta_t^i \right), \]

(33)

\[ (V_t)^{-1} = (V_t)^{-1} + (V_t)^{-1}. \]

B. Learning: The Maximization Step

The estimation of the dynamic parameters is carried out by taking the derivative of the auxiliary function (14) with respect \( \phi \). Given the formulas derived in the previous section, the terms of the auxiliary function that depend on \( \phi \) are:

\[ Q(\phi) = \frac{1}{2} \sum_{t \geq 2} \sum_{z} \rho_{t,z} \int \mathcal{N}(x_t; \eta_t, V_t) \]

(34)

\[ \times \left( \log |Q_1^{-1}| - (\eta_t - \mathbf{C}_z \eta_{t-1})^T Q_1^{-1} (\eta_t - \mathbf{C}_z \eta_{t-1}) \right) + \sum_{i=2}^{T} \sum_{i \neq j} \rho_{t,i} \log \tau_{i,j}. \]

(35)

Taking the expectation with respect to all probabilities, including the discrete state variables \( Z_t \), and using the dynamic models, we obtain:

\[ Q(\phi) = \frac{1}{2} \sum_{t \geq 2} \sum_{z} \rho_{t,z} \left( \log |Q_1^{-1}| - (\eta_t - \mathbf{C}_z \eta_{t-1})^T Q_1^{-1} (\eta_t - \mathbf{C}_z \eta_{t-1}) \right) \]

(36)

\[ - \left( (\eta_t - \mathbf{C}_z \eta_{t-1})^T Q_1^{-1} (\eta_t - \mathbf{C}_z \eta_{t-1}) \right) \]

(37)

\[ + \sum_{i=2}^{T} \sum_{i \neq j} \rho_{t,i} \log \tau_{i,j}. \]

(38)

The optimal values of the transition parameters correspond to a standard HMM model and are given by:

\[ \tau_{i,j} = \frac{1}{T-1} \sum_{t \geq 2} \rho_{t,i} \rho_{t,j}. \]

(39)

The optimal value of \( Q_z \) is obtained by taking the derivative of \( Q \) with respect to \( Q_z \) and setting this derivative equal to zero:

\[ Q_z = \frac{1}{\sum_{t \geq 2} \rho_{t,z}} \sum_{t \geq 2} \rho_{t,z} \left( (\eta_t - \mathbf{C}_z \eta_{t-1}) (\eta_t - \mathbf{C}_z \eta_{t-1})^T \right) \]

(40)

\[ + \mathbf{C}_z V_{t-1} \mathbf{C}_z^T + V_t - 2 \mathbf{C}_z W_t. \]
Taking the derivative of $Q$ with respect to $C_z$ is more involved since one needs to take the derivative of the matrix-trace operator. After setting the derivatives equal to zero we obtain:

$$C_z = \sum_{t \geq 2} \rho_t (W_t^T + \eta_t \eta_{t-1}^T) \left( \sum_{t \geq 2} \rho_t (V_{t-1} + \eta_{t-1} \eta_{t-1}^T) \right)^{-1}$$  

(36)

V. GPB2 INFERENCE AND LEARNING

GPB2 [36] is a commonly used algorithm to deal with the intractability of switching LDSs. For the sake of completeness, we describe the GPB2 algorithm for the particular case of P-LDS. As above, only the dynamic parameters $\phi$, i.e. (9), need to be estimated. The central idea of GPB2 is to recursively collapse a $K^2$-component mixture into a $K$-component mixture, as explained below. This implies that at each time index $t$ the conditional posterior $p(x_t|y_{1:t})$ is successively approximated. To do that, the marginalization chain in (10) is truncated, yielding:

$$p(x_t|y_{1:t}) = \sum_{z_{t-1}=1}^{K} \sum_{z_t=1}^{K} \int_p(x_{t-1}, x_t, z_{t-1}, z_t | y_{1:t}) dx_{t-1},$$  

(37)

with:

$$p(x_{t-1}, x_t, z_{t-1}, z_t | y_{1:t}) \propto$$  

(38)

$$p(y_t | x_t, z_t)p(x_t | x_{t-1}, z_t)p(z_t | z_{t-1})p(x_{t-1}, z_{t-1} | y_{1:t-1}).$$

GPB2 yields the following approximation:

$$p(x_t, z_{t-1} | y_{1:t-1}) = \tilde{p}_t^{*1, z_{t-1}, z_t} N(x_t | \tilde{\eta}_t^{*1, z_{t-1}, z_t}, \tilde{V}_t^{*1, z_{t-1}, z_t}),$$  

(39)

where $\tilde{p}_t^{*1, z_{t-1}, z_t}$, $\tilde{\eta}_t^{*1, z_{t-1}, z_t}$ and $\tilde{V}_t^{*1, z_{t-1}, z_t}$ play the same role as $\tilde{p}_t, \tilde{\eta}_t$ and $\tilde{V}_t$ in the previous section, but they take different numerical values because they are computed using the GPB2 approximation instead of the variational one.

By using (3), (4), (7) and (38), the approximation in (39) boils down to the following relationship:

$$p(x_t | y_{1:t}) \approx \sum_{z_{t-1}=1}^{K} \sum_{z_t=1}^{K} \tilde{p}_t^{*1, z_{t-1}, z_t} N(x_t | \tilde{\eta}_t^{*1, z_{t-1}, z_t}, \tilde{V}_t^{*1, z_{t-1}, z_t}),$$  

(40)

where the priors $\tilde{p}_t^{*1, z_{t-1}, z_t}$, the means $\tilde{\eta}_t^{*1, z_{t-1}, z_t}$, and the covariances $\tilde{V}_t^{*1, z_{t-1}, z_t}$ are given by:

$$\tilde{p}_t^{*1, z_{t-1}, z_t} = \tilde{p}_t^{*1, z_{t-1}, z_t} N(d_d, z_{t-1}, z_t),$$  

(41)

$$\tilde{\eta}_t^{*1, z_{t-1}, z_t} = \tilde{V}_t^{*1, z_{t-1}, z_t} \times \left( A_{z_t, z_{t-1}} (y_t - b_{z_t}) + P_{z_t, z_{t-1}} C_{z_t} \tilde{\eta}_t^{*1, z_{t-1}, z_t} \right)$$  

(42)

$$\tilde{V}_t^{*1, z_{t-1}, z_t} = \left( A_{z_t, z_{t-1}} \Sigma_{z_t}^{-1} A_{z_t} + P_{z_t, z_{t-1}} \Sigma_{z_t}^{-1} \right)^{-1},$$  

(43)

with:

$$d_d, z_{t-1}, z_t = y_t - A_{z_t}(C_{z_t} \tilde{\eta}_t^{*1, z_{t-1}, z_t}) - b_{z_t},$$

$$S_{z_t, z_{t-1}, z_t} = \Sigma_{z_t} + A_{z_t}(Q_{z_t} + C_{z_t} \tilde{V}_t^{*1, z_{t-1}, z_t} C_{z_t}^T)A_{z_t}^T,$$

$$P_{z_t, z_{t-1}, z_t} = (Q_{z_t} + C_{z_t} \tilde{V}_t^{*1, z_{t-1}, z_t} C_{z_t}^T)^{-1}.$$

Consequently, the dynamic model expands the $K$-component GMM hypothesized in (39) into a $K^2$-component GMM. GPB2 collapses this $K^2$ components into $K$ components by moment matching, thus obtaining:

$$p(x_t | y_{1:t}) \approx \sum_{z_{t-1}=1}^{K} \tilde{p}_t^{*1, z_{t-1}, z_t} N(x_t | \tilde{\eta}_t^{*1, z_{t-1}, z_t}, \tilde{V}_t^{*1, z_{t-1}, z_t}),$$  

(44)

where $\tilde{p}_t^{*1, z_{t-1}, z_t}, \tilde{\eta}_t^{*1, z_{t-1}, z_t}, \tilde{V}_t^{*1, z_{t-1}, z_t}$ are given by:

$$\tilde{p}_t^{*1, z_{t-1}, z_t} = \sum_{z_{t-1}=1}^{K} \tilde{p}_t^{*1, z_{t-1}, z_t},$$  

(45)

$$\tilde{\eta}_t^{*1, z_{t-1}, z_t} = \sum_{z_{t-1}=1}^{K} \tilde{\eta}_t^{*1, z_{t-1}, z_t},$$  

(46)

$$\tilde{V}_t^{*1, z_{t-1}, z_t} = \sum_{z_{t-1}=1}^{K} \tilde{V}_t^{*1, z_{t-1}, z_t}$$

(47)

Therefore, at each time index $t$, the filtering distribution is approximated with a Gaussian mixture with $K$ components, i.e. (44). The same moment matching technique can be recursively applied to the backward (or smoothing) distribution, thus obtaining a $K$-component Gaussian mixture model for $p(x_t | y_{1:T})$. Finally, it is straightforward to apply the moment matching technique to approximate the joint posterior distribution $p(x_t, x_{t-1} | y_{1:T})$ with a Gaussian mixture with $K$ components, so that the estimation of the transition parameters $\phi$ is done with the same update formulas as in the variational case, see (34), (35) and (36), but using the posterior distribution provided by GPB2.

VI. EXPERIMENTAL VALIDATION

In this section we present experimental evaluations of the proposed P-LDS variational EM filtering and smoothing algorithms, namely head pose tracking (HPT) from a video, e.g. Figure 2. The observed data consist of high-dimensional feature vectors, e.g. histogram of oriented gradients (HOG), but any other visual descriptors could be used in practice. As already explained in Section III, the static parameters can be estimated offline, from a training set of high-dimensional feature vectors (inputs) and the associated ground-truth headpose angles (outputs). In order to choose the number $K$ of linear-Gaussian components, which is equivalent to the number of states of the switch variable, we use the result of [23] which shows that the Bayesian information criterion (BIC) model [22] may be replaced with an empiric score based on the mean absolute error (MAE) between the predicted head pose and the ground truth. Based on this, we set $K = 25$ in
all our experiments. It is worthwhile to notice that the off-line training procedure is shared by the variational filter, the variational smoother and GPB2.

A. Experimental Setup

We empirically evaluate the performance of the proposed methods with three publicly available datasets: Biwi-Kinect [62], EYEDIAP [63] and Vernissage [64] (see Table I):

- The Biwi-Kinect dataset contains 24 videos of 20 people (16 men and 4 women) recorded with a Kinect camera. During the recordings, people were asked to move their heads freely in front of the camera. 3D head pose (pitch, yaw, and roll angles) annotations were obtained automatically and accurately for each video frame using the face-shift software. The angle values range from $-60^\circ$ to $60^\circ$ for pitch, $-75^\circ$ to $75^\circ$ for yaw and $-20^\circ$ to $20^\circ$ for roll. The dataset provides RGB and depth images as well as the calibration matrices. The 3D nose positions are provided as well.

- The EYEDIAP dataset is intended for both eye-gaze and head-pose estimation. It contains 94 videos of 16 people recorded using different configurations, e.g., static and rotating heads. The dataset provides RGB videos, with both HD and VGA resolution, and depth videos with the associated calibration matrices. Annotations of both head-pose and eye-gaze are provided for each video frame. The angle values range from $-40^\circ$ to $40^\circ$ for pitch and $-50^\circ$ to $50^\circ$ for yaw.

- The Vernissage dataset contains 10 recordings of 20 people interacting with each and with a robot. Each recording comprises two people. The dataset was recorded with a camera embedded into the head of a robot head. A network of infrared cameras combined with optical markers placed on the participants’ heads provide accurate ground-truth head positions and orientations.

B. Implementation details

Facial features are computed as follows. A face detector provides bounding boxes for each frame and for each video. Then a high-dimensional feature vector is extracted from each bounding box using histogram of oriented gradients (HOG) descriptors, obtained along the implementation described in [23] which yields vector-valued observations of dimension $D = 1888$. For all datasets and for each face identity, we split the corresponding videos into two disjoint sets: a training set and a test set. Since the datasets are annotated, i.e. there are ground-truth head-pose parameter values associated with each frame. As already mentioned, we use the method of [23] to estimate $\theta$.

The dynamic parameters are initialized as follows. First we set $C_k = 1$ since we noticed that the simultaneous estimation of $C_k$ and $Q_k$ is subject to instabilities in the estimation of the dynamic parameters. The covariance matrices $\{C_j, Q_j\}_{j=1}^K$ are initialized with the identity matrix. The entries of the transition matrix $\{\tau_{ij}\}_{i,j=1}^K$ are initialized in the following way. We compute the pairwise Bhattacharrya distances [65] between the $K$ Gaussian components defined by (5). The variational EM algorithm alternates between inference of $q(Z_k)$ (E-Z step) and of $q(x_k)$ (E-X step). The variational parameters $\eta_t$ and $V_t$ are initialized with their previously estimated values, namely: $\eta_t = \eta_{t-1}$ and $V_t = V_{t-1}$.

We implemented both the variational filter and the variational smoother described in detail in the previous section. The main difference between the filter and the smoother is the amount of information that is used at inference time. Indeed, while the filter uses only causal information (i.e. past observations) the smoother uses also non-causal information (i.e. past and future observations). This is why, a priori, one expects the smoother to have better performance, at the price of being a completely off-line algorithm that cannot be used for real-time applications.

C. Results and Discussion

The proposed methods were compared with the following state-of-the-art HPT methods:

- **Landmarks** [66], [67] combines 2D face landmark detection with head pose estimation and with tracking. Head pose is estimated in the following way: the 3D landmarks of a mean face are projected onto the image plane and the error between the projected landmarks and the observed landmarks is minimized over the pose parameters. At each time step this non-linear minimizer is initialized with the pose parameters computed at the previous time step. The publicly available implementation of this method only computes the pitch and yaw angles.

- **OpenFace** [68] is an open source software package for facial behavior analysis, i.e. facial landmark detection and tracking, head pose and eye gaze estimation. It extracts 68 2D facial landmarks using conditional local neural fields and tracks them over time with a three-layer CNN trained to predict landmark detection errors [69]. Once 2D landmarks are detected and tracked, they are used in conjunction with a 3D facial model to compute head pose parameters.

- **ICP tracking** [70] uses both depth and color (RGB-D images) from which 16 3D facial landmarks are manually estimated. The distance errors are computed using the ICP algorithm.

---

### Table I

**Summary of principal features of the datasets used for empirical evaluation.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#Recordings</td>
<td>24</td>
<td>94</td>
<td>10</td>
</tr>
<tr>
<td>#Participants</td>
<td>20</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>Sensor type</td>
<td>RGB-D</td>
<td>RGB-D</td>
<td>RGB</td>
</tr>
<tr>
<td>Pitch range</td>
<td>$[-60^\circ, +60^\circ]$</td>
<td>$[-40^\circ, +40^\circ]$</td>
<td>$[-90^\circ, +90^\circ]$</td>
</tr>
<tr>
<td>Yaw range</td>
<td>$[-75^\circ, +75^\circ]$</td>
<td>$[-50^\circ, +50^\circ]$</td>
<td>$[-90^\circ, +90^\circ]$</td>
</tr>
<tr>
<td>Roll range</td>
<td>$[-20^\circ, +20^\circ]$</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Annotation</td>
<td>Automatic fitting with a deformable 3D shape model</td>
<td>Optical motion capture device</td>
<td></td>
</tr>
<tr>
<td>Sensor type</td>
<td>RGB-D</td>
<td>RGB-D</td>
<td>RGB</td>
</tr>
<tr>
<td>Pitch range</td>
<td>$[-60^\circ, +60^\circ]$</td>
<td>$[-40^\circ, +40^\circ]$</td>
<td>$[-90^\circ, +90^\circ]$</td>
</tr>
<tr>
<td>Yaw range</td>
<td>$[-75^\circ, +75^\circ]$</td>
<td>$[-50^\circ, +50^\circ]$</td>
<td>$[-90^\circ, +90^\circ]$</td>
</tr>
<tr>
<td>Roll range</td>
<td>$[-20^\circ, +20^\circ]$</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Annotation</td>
<td>Automatic fitting with a deformable 3D shape model</td>
<td>Optical motion capture device</td>
<td></td>
</tr>
</tbody>
</table>
extracted. The landmarks are tracked based on estimating
the rigid motion between consecutive frames.

These HPT methods make use of 2D facial landmarks, whose detection is known to be sensitive to large head
poses that induce partial occlusions of the face. This stays
in contrast with the proposed method that directly exploits
high-dimensional descriptors of faces. For completeness, we
also compared our algorithms with an implementation of the
Kalman filter (LDS), namely (1)-(6) with \( K = 1 \), with the
HPE method of [23] and with GPB2, i.e. Section V and [71].
To quantitatively evaluate HPT we compute average and
standard deviation of the absolute error between the estimated
pose parameters and the ground-truth parameters provided
with each annotated dataset.

The results obtained with the Biwi-Kinect, EYEDIAP and
Vernissage datasets are summarized in tables II, III and IV.
The HPT method of [66], [67] does not estimate the roll
angle. Moreover, this method relies on 2D landmark detection.
Therefore, it fails whenever all the landmarks are not properly
detected. This is the reason for which this method did not
provide publishable results, when applied to the EYEDIAP and
Vernissage datasets. The method of [70] takes as input RGB-
D images and manually annotated facial landmarks, hence it
could only be applied to the EYEDIAP dataset. In general,
the proposed descriptor-based trackers perform better than the
landmark-based trackers. A lower standard deviation corre-
sponds to a higher precision and a better method repeatability.

In order to assess the statistical differences in performance
between the various methods, we used the Wilcoxon signed-
rank test [72], inspired from [73], which is a non-parametric
statistical hypothesis test that is commonly used to compare
two related samples in order to assess the null hypothesis that
the median difference between pairs of observations is zero. It
can be used as an alternative to the paired Student’s t-test (t-
test for matched pairs) when the population cannot be assumed
to be normally distributed. For most of the comparisons, there
were no statistical differences. The only notable exception is
OpenFace [68], which performs considerably worse than the
other methods on two of the three datasets: the yaw and roll
estimation for the Biwi-Kinect dataset, and the pitch and yaw
estimation for the EYEDIAP dataset. These estimation
are marked with an asterisk in Table II and Table III, respectively.

As an example, Figure 3 shows a ground-truth yaw (left-
right rotation) trajectory as well as trajectories obtained with HPE_GLLiM [23], with HPT_GPB2 (top) and with the proposed variational filter (bottom). Overall, the performance of the proposed variational EM algorithms is comparable with the performance of GPB2. Notice that the large errors that produced by HPE_GLLiM are eliminated by both trackers.

We measured the CPU time needed to compute one time step. This comprises the following processes: (i) extraction of a HOG descriptor from a face, (ii) head pose estimation, and (iii) tracking, where the tracker can be either the GPB2 or the variational filter. Using an Intel-Xeon and Matlab, it takes 1.0 second to extract a HOG descriptor and to estimated head pose, 8.55 seconds to run GPB2 and 2.45 seconds to run the variational filter, respectively.

VII. Conclusions

In this paper we addressed the problem of learning and inference of piecewise LDS. The latter belongs to the switching LDS class of models, which is known to be intractable because of the combinatorial explosion, over time, of the modes of the latent space. The standard way of dealing with this problem is to use the GPB2 algorithm which collapses a $K^2$-component mixture into a $K$-component one based on moment matching – a computationally demanding process.

Alternatively, we propose a variational approximation of P-LDS and two associated EM algorithms: a variational filter and a variational smoother. Both the filter and the smoother are based on closed-form expressions, which guarantees efficient computation and fast convergence. The proposed variational filter is of the order of 3.5 times faster than the GPB2 method. Not surprisingly, the most time-consuming part of GPB2 is the computation and fast convergence. The proposed variational filter is of the order of 3.5 times faster than the GPB2 method. Not surprisingly, the most time-consuming part of GPB2 is the evaluation of the parameters of the $K^2$-component Gaussian mixture, followed by the evaluation of the parameters of the approximating $K$-component mixture. This collapsing process resides at the heart of GPB2 and it cannot be avoided. In contrast, the M-step of the proposed VEM can be skipped, once the parameters are learnt, as done in [2], which can further accelerate the algorithm.

We applied the proposed algorithms to the problem of headpose tracking. We presented a series of experiments using several datasets. We carried out a benchmark that included our algorithms and several state-of-the-art tracking algorithms. We note that the variational-based tracker compares well with GPB2 and with the other trackers. It should be noted, however,

**Table III**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HPE_GLLiM_HOG [23]</td>
<td>6.29</td>
<td>7.80</td>
<td>7.80</td>
<td>10.39</td>
</tr>
<tr>
<td>ICP tracking [70]</td>
<td>4.17</td>
<td>5.59</td>
<td>6.89</td>
<td>14.42</td>
</tr>
<tr>
<td>OpenFace [68]</td>
<td>15.39*</td>
<td>12.85</td>
<td>22.21*</td>
<td>16.32</td>
</tr>
<tr>
<td>HPT_GLLiM_Kalman</td>
<td>6.21</td>
<td>7.75</td>
<td>10.62</td>
<td>10.31</td>
</tr>
<tr>
<td>HPT_GPB2_HOG</td>
<td>6.68</td>
<td>8.75</td>
<td>8.44</td>
<td>10.91</td>
</tr>
<tr>
<td>HPT_VarFilter_HOG</td>
<td>6.96</td>
<td>8.04</td>
<td>11.38</td>
<td>11.44</td>
</tr>
<tr>
<td>HPT_VarSmother_HOG</td>
<td>6.78</td>
<td>7.88</td>
<td>10.66</td>
<td>10.99</td>
</tr>
</tbody>
</table>

**Table IV**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HPE_GLLiM_HOG [23]</td>
<td>23.95</td>
<td>23.18</td>
<td>11.03</td>
<td>8.57</td>
</tr>
<tr>
<td>OpenFace [68]</td>
<td>21.30</td>
<td>18.80</td>
<td>13.18</td>
<td>10.67</td>
</tr>
<tr>
<td>HPT_GLLiM_Kalman</td>
<td>23.94</td>
<td>23.18</td>
<td>11.03</td>
<td>8.56</td>
</tr>
<tr>
<td>HPT_GPB2_HOG</td>
<td>20.24</td>
<td>20.62</td>
<td>10.21</td>
<td>7.80</td>
</tr>
<tr>
<td>HPT_VarFilter_HOG</td>
<td>21.06</td>
<td>19.96</td>
<td>13.76</td>
<td>8.25</td>
</tr>
<tr>
<td>HPT_VarSmother_HOG</td>
<td>20.37</td>
<td>18.58</td>
<td>12.92</td>
<td>7.89</td>
</tr>
</tbody>
</table>
that the landmark based methods, e.g. [66], [67], fail to track in many cases. The best performing method is [70]. Nevertheless, this method has limitations because it requires manual annotation of facial landmarks and it can only be applied to RGB-D images. In contrast, the proposed method is based on extracting high-dimensional descriptors from RGB images, and appears to be more robust against facial self occlusions than the landmark-based methods.

REFERENCES


Xavier Alameda-Pineda received M.Sc. degrees in mathematics (2008), in telecommunications (2009) and in computer science (2010) and a Ph.D. in mathematics and computer science (2013) from Université Joseph Fourier. Since 2016, he is a research scientist at Inria Grenoble Rhône-Alpes, with the Perception team. He served as Area Chair at ICCV’17, of ICIAP’19 and of ACM MM’19. He is the recipient of several paper awards and of the ACM SIGMM Rising Star Award in 2018. He is the coordinator of the I2020 ICT SPRING project.

His scientific interests lie in computer vision, machine learning and signal processing for robotics and multimodal social behavior analysis.

Vincent Drouard received the M.Sc. in Applied Mathematics from the School of Engineering, Université Nice Sophia-Antipolis in 2013 and the Ph.D. in Computer Science from Université Grenoble Alpes in 2017. During the period 2014-2017 he was a member of the Perception team at Inria Grenoble Rhône-Alpes. Currently he is a research scientist with Image Metrics, Manchester, UK. He received the best student paper award at IEEE International Conference on Image Processing (ICIP’15).

Vincent’s scientific interests lie in computer vision, machine learning, and signal processing. He is a member of the Perception team at Inria Grenoble Rhône-Alpes. His research interests include computer vision, machine learning, and signal processing. He holds a Ph.D. in Computer Science from Université Grenoble Alpes and is a member of the Perception team at Inria Grenoble Rhône-Alpes. His research interests include computer vision, machine learning, and signal processing. He is also a member of the Perception team at Inria Grenoble Rhône-Alpes. His research interests include computer vision, machine learning, and signal processing. He is currently a research scientist with Image Metrics, Manchester, UK. He received the best student paper award at IEEE International Conference on Image Processing (ICIP’15).

Radu Horaud received the B.Sc. degree in Electrical Engineering (1977), the M.Sc. degree in Control Engineering (1979), and the Ph.D. degree in Computer Science (1981), all from Grenoble INP, France. In the past he held research positions with SRI International (1982-1984) and with CNRS (1984-1998). Since 1998 he has been director of research with Inria Grenoble Rhône-Alpes, where he leads the Perception team. His research interests include computer vision, machine learning, audio signal processing, audio-visual analysis, and robotics. Radu Horaud and his collaborators received numerous best paper awards. He was an area editor of Computer Vision and Image Understanding, an associate editor of the International Journal of Computer Vision, and an editorial board member of the International Journal of Robotics Research. Radu Horaud was awarded two ERC grants, an advanced grant for the project Vision and Hearing in Action (2014-2019) and a proof of concept grant (2018-2019).