Switching Gaussian Process Dynamic Models for Simultaneous Composite Motion Tracking and Recognition

Jixu Chen Minyoung Kim Yu Wang Qiang Ji
Department of Electrical,Computer and System Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
{chenj4,kimm8,wangy15}@rpi.edu qji@ecse.rpi.edu

Abstract

Traditional dynamical systems used for motion tracking cannot effectively handle high dimensionality of the motion states and composite dynamics. In this paper, to address both issues simultaneously, we propose the marriage of the switching dynamical system and recent Gaussian Process Dynamic Models (GPDM), yielding a new model called the switching GPDM (SGPDM). The proposed switching variables enable the SGPDM to capture diverse motion dynamics effectively, and also allow to identify the motion class (e.g. walk or run in the human motion tracking, smile or angry in the facial motion tracking), which naturally leads to the idea of simultaneous motion tracking and classification. Moreover, each of GPDMs in SGPDM can faithfully model its corresponding primitive motion, while performing tracking in the low-dimensional latent space, therefore significantly improving the tracking efficiency. The proposed SGPDM is then applied to human body motion tracking and classification, and facial motion tracking and recognition. We demonstrate the performance of our model on several composite body motion videos obtained from the CMU database, including exercises and salsa dance. We also demonstrate the robustness of our model in terms of both facial feature tracking and facial expression/pose recognition performance on real videos under diverse scenarios including pose change, low frame rate and low quality videos.

1. Introduction

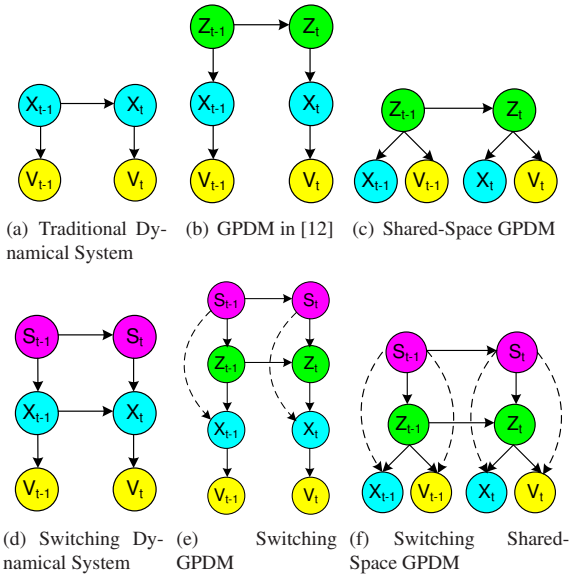
Traditional dynamical systems and their variants are widely used to tackle the tracking problems. Traditional dynamical models include the popular Kalman filtering and particle filtering as shown in Figure 1(a). Although these dynamical systems are commonly used to represent dynamics in the motion space, they have two main limitations: (1) they are not efficient in modeling complex dynamics with a large number of model parameters, and (2) they are inherently limited to modeling single monolithic motion dynamics, inappropriate to account for composite motions with diverse dynamics.

In the community, the two issues of high dimensionality of the motion and composite dynamics have been tackled considerably, but often individually. For instance, the switching dynamical systems [8] can resolve the latter issue, yet suffering from modeling complexity. On the other hand, the recent Gaussian Process Dynamic Models (GPDM) [5, 12] (Figure 1(b)), which are inspired by the Gaussian Process Latent Variable Model [4], can alleviate the first limitation by discovering the intrinsic low-dimensional latent representation of the model dynamics. The latent embedding, latent dynamics and reconstruction mapping can be learned simultaneously from the training sequences [12]. However, the model capacity of GPDMs may be insufficient to capture multiple heterogeneous dynamics by nature.

In this paper, we propose to combine Switching Model with the Gaussian Process Dynamic Models to produce a dynamic model called the Switching Gaussian Process Dynamic Model (SGPDM) as shown in Fig.1(e) and Fig.1(f). This model has a switching layer on top of the latent variables. Hence, it enjoys both worlds of switching dynamics and low dimensional dynamics with significantly reduced number of parameters. By incorporating a switching layer this model also allows simultaneous motion tracking and recognition. To demonstrate the proposed model, we then consider two representative composite motion tracking problems, the human body motion tracking and facial motion tracking. Human motion tracking or facial motion tracking(in this paper, facial motion tracking refers to facial feature tracking) is a task of predicting the sequence of X , which are body poses (typically 3D joint angles of human body) or facial features (typically coordinates of the

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

fiducial points of facial features) from the image measurement sequence V. The image measurement is represented by the moment features extracted from the silhouette images as shown in Fig. 2 or Gabor features around the facial feature points.



(a) Traditional Dynamical System (b) GPDM in [12] (c) Shared-Space GPDM

(d) Switching Dynamical System (e) Switching GPDM (f) Switching Shared-Space GPDM

Figure 1. Graphical representation of dynamic models for tracking. X,V,Z and S represent the state (body or facial pose), image measurement, latent variable and switching state respectively. (a) is the traditional dynamic system, such as Kalman filter and particle filter. (b) is the GPDM model used in [12]. (c) is the proposed shared-space GPDM model (Section 3.2.2). (d) is the switching dynamic system used in [8]. (e) is the Switching GPDM model, an extension of (b) (Section 4.2) (f) is the Switching Shared-Space GPDM model, an extension of (c) (Section 4.1)

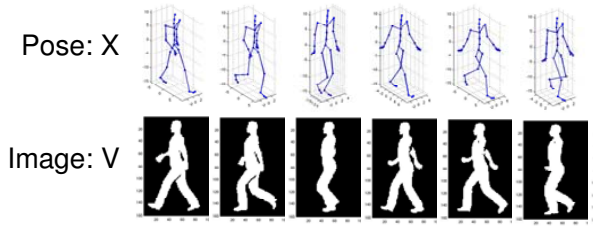


Figure 2. The state X and image measurement V for body pose tracking

The dynamic systems of Human body/facial motion have high dimensionality of X and multiple heterogeneous dynamics. For example, the body pose or facial features is typically composed of several dozens of joint angles or fiducial points which would produce too many parameters for modeling dynamics directly in the X space. Human body/facial motion also contain different styles and contents (e.g, facial feature points are subject to the pose and expression change), which may not be faithfully represented by standard dynamic models incorporating a single dynamics.

We first apply the Switching GPDM model on the human body motion estimation/tracking problem, where we predict the body pose (3D joint angles) from the observed sequence of image silhouettes. Then we consider the problem of simultaneous facial feature tracking and facial expression/pose recognition on real video, under diverse scenarios including pose change, low frame rate and low quality videos.

2. Related works

In this section, some works related to latent variable models, switching systems, and simultaneous motion tracking and recognition methods are reviewed. The high dimensionality of state space in the tracking problem can be circumvented by introducing the low dimensional latent variables (Z) on top of the state variables, and model the dynamics in this low dimensional space. For example, Urtasun et.al [12] introduced the Gaussian Process Dynamic Model (GPDM) for body tracking (as shown in Fig.1(b)). Z indicates the intrinsic manifold maps that generate the pose X. With the incorporated nonlinear Gaussian Process based dimensionality reduction methods, GPDM showed significant performance improvement in Human motion tracking.

Besides the standard GPDM model in Fig.1(b), we also introduce another form of GPDM in this paper, as shown in Fig.1(c). Similar to the standard model, it also models the dynamics in a learned low-dimensional latent space based on Gaussian Process, where the latent space is shared by both the pose (X) and the image measurement (V). We call it Shared-Space GPDM, and more details are discussed in section 3.2.2. Please notice that since Z is hidden and is not known, V and X are dependent.

However, these dimensionality reduction approaches basically assume a single dynamics, preventing it from being successfully applied to composite motions. To address the composite motion issue, the so-called switching dynamic systems have been proposed by [8] (Fig.1(d)). Here, the latent switching variables (S) effectively represent switching over different motion dynamics. Specifically, in the switching dynamic systems, the latent switching variable (S) has several discrete states. It automatically selects the current member from a set of dynamic models to increase the robustness and accuracy of the motion tracking, and also performs classification of current motion. This naturally leads to the idea of simultaneous motion recognition and tracking.

Tracking and recognition have been often tackled individually as independent task. For example most existing methods for facial expression recognition generally involve two steps, tracking facial features is followed by the expression recognition based on extracted facial features.

Simultaneous tracking and recognition which aims to exploit the synergy of tracking and recognition is not a new idea and some efforts for simultaneous tracking and recognition [7, 15, 1] already exist in the literature. In [7], a simultaneous tracking and recognition method based on mixed-state CONDENSATION [3] is designed for tracking and recognizing the juggled balls. Zhou et al. [15] proposes a probabilistic framework allowing simultaneous tracking and recognition of Human face from video. In [1], a method is presented for simultaneous facial action tracking and expression recognition. In this method, there are three layers in the graphical model. The first layer represents the expression states which decide the dynamic model of the facial action control vector of the employed 3D face model in the second layer. The facial action control vector along with the face pose parameters recovered from a separate deterministic optimization framework are used for warping the face into geometrically free facial patch, which is then compared with an on-line appearance model to output an observation likelihood. Inference of this model is also based on the mixed-state CONDENSATION.

However, like other switching systems, existing simultaneous methods based on particle filter either are limited to tracking a motion of limited dimension because the workable dimensionality for particle filter is small (e.g. the facial action control vector in [1] involves only the eyebrow and mouth of the face, and has a dimension of six), or employ weak (linear) dynamic models (e.g. the auto-regressive (AR) model is used as the transitional model of facial action in [1]).

Compared to the existing simultaneous methods in the literature, by combining GPDM with switching variables, our Switching GPDM model for simultaneous motion tracking and recognition is capable of tracking high-dimensional motion (e.g. there are 59 dimension for the joint angles in the human pose tracking model and 56 dimensions for 28 facial feature points in our simultaneous facial feature tracking model. None of the simultaneous methods mentioned above can track such high-dimensional motion), and has the power of handling more complex (non-linear) dynamics.

In summary, the contributions of this paper are listed as follows: First, a new model called Switching Gaussian Process Dynamic Model is proposed by combining GPDM and switching variables. It addressed the two issues of traditional tracking system simultaneously. Second, besides pure tracking problems, the proposed Switching GPDM can also perform simultaneous tracking and recognition. We test its capability in both tracking and recognition in the experiments.

3. Proposed Approach

3.1. Gaussian Process Latent Variable Model

Here we briefly review the Gaussian Latent Variable Model (GPLVM), which is the basis for GPDM. GPLVM proposed by Lawrence [4] is an efficient tool to model the distribution in a high dimensional space with a compact low dimension representation. It has been previously used to provide a prior probability of human pose for animation [2] and body tracking [13, 10]

3.1.1 Gaussian Process

We first start with the Gaussian Process (GP). GP is a non-parametric approach for solving regression problem, which learns a mapping $\mathbf{y} = f(\mathbf{z})$ from some training pairs $\{\mathbf{z}_i, \mathbf{y}_i\}_{i=1}^N$, where each $\mathbf{y}_i \in \mathbb{R}^D$, and $\mathbf{z}_i \in \mathbb{R}^S$. Arrange the training vectors into the rows of matrices $\mathbf{Y}^{N\times D} = [\mathbf{y}_1, ..., \mathbf{y}_N]^T$ and $\mathbf{Z}^{N\times S} = [\mathbf{z}_1, ..., \mathbf{z}_N]^T$. The conditional probability of $\mathbf{Y}$ given $\mathbf{Z}$ is defined based on GP:

$$p(\mathbf{Y}|\mathbf{Z}, \beta) = \prod_{d=1}^{D} N(\mathbf{Y}_{:,d}|0, \mathbf{K(Z,Z)}) \tag{1}$$

$\mathbf{Y}_{:,d}$ indicate the d^{th} column of $\mathbf{Y}$, which is a $N \times 1$ vector constructed from the d^{th} dimension of the training data. **K(Z,Z)** is a $N \times N$ covariance matrix whose entries are given by the kernel function:

$$k(\mathbf{z}_i, \mathbf{z}_j) = \exp(-\beta\|\mathbf{z}_i - \mathbf{z}_j\|^2) \tag{2}$$

The learning process of GP is to adjust the parameter β to maximize the probability in Eq. 1.

Once the parameters have been learned, the GP prediction of $\mathbf{y}$, given a new input $\mathbf{z}$, can be derived as a Gaussian distribution (For full details, please refer to [4] or [9]):

$$p(\mathbf{y}|\mathbf{z}) = p(\mathbf{y}|\mathbf{z}, \mathbf{Y}, \mathbf{Z}, \beta) = N(\mu(\mathbf{z}), \sigma^2(\mathbf{z})\mathbf{I}) \tag{3}$$

with:

$$\begin{aligned} \mu(\mathbf{z}) &= \mathbf{Y}^T\mathbf{K(Z,Z)}^{-1}\mathbf{k(z,Z)} \\ \sigma^2(\mathbf{z}) &= k\mathbf{(z,z)} - \mathbf{k(z,Z)}^T\mathbf{K(Z,Z)}^{-1}\mathbf{k(z,Z)} \end{aligned} \tag{4}$$

where **k(z, Z)** is a $N \times 1$ vector whose i^{th} entry is $k(\mathbf{z}, \mathbf{z}_i)$.

Notice that, the covariance matrix of this Gaussian distribution is diagonal, which means given $\mathbf{z}$, each dimension of $\mathbf{y}$ is conditionally independent with others.

3.1.2 GPLVM

Instead of knowing the complete training pairs $\{\mathbf{z}_i, \mathbf{y}_i\}_{i=1}^N$, GPLVM learning is an unsupervised process where we are

given only $\mathbf{Y}$, and we need to optimize the kernel functions and the latent variable $\mathbf{Z}$ together:

$$\mathbf{Z}^*, \beta^* = \arg\max_{\mathbf{Z},\beta} p(\mathbf{Y}|\mathbf{Z},\beta)p(\mathbf{Z}) \quad (5)$$

where the first term is from Eq. 1 and the second term, i.e., the prior of the latent variable, is defined as Gaussian [2, 13, 10]:

$$p(\mathbf{Z}) = \prod_{i=1}^{N} N(\mathbf{z}_i|0,\mathbf{I}) \quad (6)$$

It is known that this optimization is non-convex, and it does not admit a closed-form solution. The scaled conjugate gradient (SCG) method search is often used, and known to be effective [4]. We use the SCG for optimization where the latent positions $\mathbf{Z}$ are initialized as the PCA coefficients of $\mathbf{Y}$

Usually, the latent space $\mathbb{Z}$ has much lower dimension than $\mathbb{Y}$. For example, in human pose estimation, the pose $\mathbf{y}$ usually have 20-50 dimensions, while the latent variable $\mathbf{z}$ only need 2-3 dimensions, which is a widely believed manifold dimensionality for single motions like walking and running. So, once GPLVM is learned, the pose can be represented by a low-dimensional variable $\mathbf{z}$, and the mapping from $\mathbf{z}$ to $\mathbf{y}$ ($p(\mathbf{y}|\mathbf{z})$) is modeled as Eq. 3.

3.2. Gaussian Process Dynamic Model

The basic idea of using GPLVM for tracking is modeling the dynamics in the low-dimensional latent space instead of modeling it in the high-dimensional pose space. Two typical latent variable tracking models are shown in Fig. 1(b) and Fig. 1(c).

3.2.1 GPDM

Fig. 1(b) shows the standard GPDM proposed in [12]. $\mathbf{x}_t \in \mathbb{R}^D$ represent the current body pose, $\mathbf{z}_t \in \mathbb{R}^S$ represent the corresponding low-dimensional latent variable.

Same as GPLVM, given the training sequence: $\mathbf{X}^{T\times D} = [\mathbf{x}_1, ..., \mathbf{x}_T]^T$, the mapping from low-dimensional latent space to pose space is modeled by GP regression:

$$p(\mathbf{X}|\mathbf{Z},\beta_X) = \prod_{d=1}^{D} N(\mathbf{X}_{:,d}|0,\mathbf{K}_X(\mathbf{Z},\mathbf{Z})) \quad (7)$$

where β_X is the kernel parameter for this mapping GP regression. Taking one step further, the dynamics in the latent space is also modeled by GP regression:

$$p(\mathbf{Z}^+|\mathbf{Z}^-,\beta_T) = \prod_{s=1}^{S} N(\mathbf{Z}^+_{:,s}|0,\mathbf{K}_T(\mathbf{Z}^-,\mathbf{Z}^-)) \quad (8)$$

where β_T is the kernel parameter for the dynamic GP regression, and $\mathbf{Z}^+ = [\mathbf{z}_2, ..., \mathbf{z}_T]^T$, $\mathbf{Z}^- = [\mathbf{z}_1, ..., \mathbf{z}_{T-1}]^T$. The GPDM is then trained to maximize:

$$\mathbf{Z}^*, \beta_X^*, \beta_T^* = \arg\max_{\mathbf{Z},\beta_X,\beta_T} p(\mathbf{X}|\mathbf{Z},\beta_X)p(\mathbf{Z}^+|\mathbf{Z}^-,\beta_T) \quad (9)$$

Once the model is learned, following Eq.3, the conditional probabilities needed for tracking can be written as Gaussian distribution: $p(\mathbf{x}_t|\mathbf{z}_t) = p(\mathbf{x}_t|\mathbf{z}_t,\mathbf{X},\mathbf{Z},\beta_X)$ and $p(\mathbf{z}_t|\mathbf{z}_{t-1}) = p(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{Z}^+,\mathbf{Z}^-,\beta_T)$.

Notice that, for this standard GPDM tracking model, the relationship between pose $\mathbf{x}_t$ and measurement $\mathbf{v}_t$ ($p(\mathbf{v}_t|\mathbf{x}_t)$, the link from $\mathbf{x}_t$ to $\mathbf{v}_t$ in Fig.1(b)) is not modeled by Gaussian Process. In [12], this likelihood of pose is defined based on the detected joint points. In our face tracking algorithm, the likelihood is defined based on Gabor feature matching (Section 4.2).

3.2.2 Shared Space GPDM

However, in our body tracking experiment (Section 4.1), the measurement is a moment feature from silhouette. It is difficult to directly model the relationships between pose and the moment feature. Although some other learning based method, such as neural network [10], has been used to learn this relationship, we propose to model it also though Gaussian Process for consistency. Fig. 1(c) shows our GPDM for body tracking. The pose $\mathbf{v}_t$ and measurement $\mathbf{x}_t$ are combined as the training data:

$$\mathbf{y}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{v}_t \end{pmatrix} \quad (10)$$

Then we learn the latent space from the training sequence $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_T]^T$. The learning process is same as the standard GPDM. However, because Y includes both pose and measurement, we call the learned latent Z space "shared latent space". (More detailed discussion of shared latent space using GPLVM can be found in [6]). Similarly, once the GPDM is learned, the conditional probability can be written as Gaussian distributions $p(\mathbf{y}_t|\mathbf{z}_t) = p(\mathbf{y}_t|\mathbf{z}_t,\mathbf{Y},\mathbf{Z},\beta_Y)$ and $p(\mathbf{z}_t|\mathbf{z}_{t-1}) = p(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{Z}^+,\mathbf{Z}^-,\beta_T)$ (Eq.3). Notice that in Eq.3, the covariance matrix of this Gaussian is diagonal matrix, so the conditional probability can be factorized as Eq.11, and $\mathbf{y}_t$ can be represented by two separate nodes: $\mathbf{x}_t$ and $\mathbf{v}_t$ connected to $\mathbf{z}_t$ in Fig.1(c).

$$p(\mathbf{y}_t|\mathbf{z}_t) = p(\begin{pmatrix} \mathbf{x}_t \\ \mathbf{v}_t \end{pmatrix}|\mathbf{z}_t) = p(\mathbf{x}_t|\mathbf{z}_t)p(\mathbf{v}_t|\mathbf{z}_t) \quad (11)$$

3.3. Switching Gaussian Process Dynamic Model

The above GPDMs basically assume a single dynamics in the latent $\mathbb{Z}$ space. In order to deal with the composite motion dynamics, we extend the GPDM to the Switching GPDMs whose graphical representations are shown in

Fig.1(e) and Fig.1(f). The three layers of a Switching GPDM are denoted in Fig.3, which is an extension of shared space GPDM. We also extended the standard GPDM as shown in Fig.1(e). But due to the similarity of two models, we will only discuss the model in Fig.3.

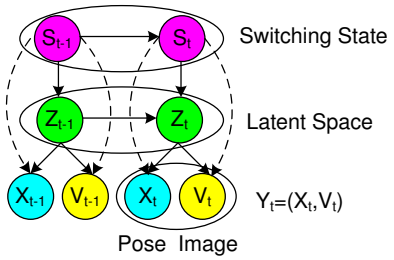


Figure 3. The three layers of a Switching GPDM

Roughly, a Switching GPDM is composed of the dynamics of the switching component $p(s_t|s_{t-1})$, and for each switching state $s_t = c$, we have a specific GPDM(motion dynamic model) denoted by GPDM^c,where $c = 1, ..., M$ (Assuming M different dynamics).

3.3.1 Learning

We first discuss how to train the Switching GPDM model in the supervised case, where we are given the true switching state values $\mathbf{S} = [s_1, ..., s_T]$. Notice that the learning criterion is the joint log-likelihood:

$$\log p(\mathbf{S}, \mathbf{Z}, \mathbf{Y}) = \log p(\mathbf{S}) + \log p(\mathbf{Z}|\mathbf{S}) + \log p(\mathbf{Y}|\mathbf{S},\mathbf{Z}). \tag{12}$$

Knowing the switching variables makes the objective decomposed along the switching values, and we independently learn the individual GPDM models. We need to estimate the parameters $(\theta, \lambda^{(1..M)})$, where θ denotes the dynamics of the switching state, and $\lambda^{(c)} = (\mathbf{Z}^{(c)}, \beta_Y^{(c)}, \beta_T^{(c)})$ denotes the parameters of the c^{th} GPDM, $c = 1..M$.

For example, given a 10-frame labeled sequence where $t = 1..4$ and $t = 8..10$ are labeled as the 1^{st} GPDM, and the rest frames are labeled as the 2^{nd} GPDM. Then, we can maximize the following log-likelihood:

$$\begin{aligned}&\log p(\mathbf{S}, \mathbf{Z}, \mathbf{Y}|\theta, \lambda^{(1..2)}) = \\ &\log p(s_{1..10}|\theta) + \log p(s_{1..4,8..10}, \mathbf{z}_{1..4,8..10}, \mathbf{y}_{1..4,8..10}|\lambda^{(1)}) \\ &+ \log p(s_{5..7}, \mathbf{z}_{5..7}, \mathbf{y}_{5..7}|\lambda^{(2)})\end{aligned} \tag{13}$$

While training individual GPDMs, we should take care of some boundary slices(e.g., t=4 and t=7), which commonly appears in two GPDMs.

For the unsupervised case, where the switching states are not labeled in the training sequence, we optimize $p(\mathbf{S}, \mathbf{Z}, \mathbf{Y})$ with respect to $\mathbf{S}$ and $\mathbf{Z}$ iteratively. In each iteration, we first estimate the most likely $\mathbf{S}$ given the current model, then the model is updated using the estimated $\mathbf{S}$ through the supervised learning algorithm.

3.3.2 Tracking

Tracking in the SGPDM is the task to compute $p(\mathbf{x}_t|\mathbf{v}_1, ..., \mathbf{v}_t)$ or to find its mode or mean. The two hidden layers (s_t and $\mathbf{z}_t$) can be merged into a single layer, namely by introducing the so-called super-node $\mathbf{u}_t = (s_t, \mathbf{z}_t)$, and we can readily run the standard particle filtering algorithm. More specifically, we assume the weighted samples (particles) are given at time $t-1$, namely,

$$p(\mathbf{u}_{t-1}|\mathbf{v}_1, ..., \mathbf{v}_{t-1}) \approx \{u_{t-1}^{(i)}, w_{t-1}^{(i)}\}_{i=1}^n \tag{14}$$

Then, we estimate the particles at time t by the following steps:

1. Re-sample from $\{u_{t-1}^{(i)}, w_{t-1}^{(i)}\}_{i=1}^n$
2. Let the samples from step.1 undergo the dynamics $p(\mathbf{u}_t|\mathbf{u}_{t-1})$, which is done by first propagating the switching state $p(s_t|s_{t-1})$ followed by the Gaussian Process dynamics $p(\mathbf{z}_t|\mathbf{z}_{t-1}, s_t)$.
3. Re-weight the samples $\mathbf{u}_t$ according to the GP likelihood $p(\mathbf{v}_t|\mathbf{u}_t) = p(\mathbf{v}_t|\mathbf{z}_t, s_t)$.
4. Numerically integrate the samples $\mathbf{u}_t$ and the weights from step.3, as follows:

$$p(\mathbf{x}_t|\mathbf{v}_1, ..., \mathbf{v}_t) = \int_{u_t} p(\mathbf{x}_t|\mathbf{u}_t)p(\mathbf{u}_t|\mathbf{v}_1, ..., \mathbf{v}_t) \tag{15}$$

4. Experiment

To validate the performance of the proposed SGPDM, we applied it to two different applications: human body motion tracking and classification, and facial feature tracking and expression/pose recognition.

4.1. Human Body Tracking

We consider 3D motion capture data from the CMU MoCap database (http://mocap.cs.cmu.edu). The pose $\mathbf{x}$ is composed of 59 joint angles of human body. We study the performance of Switching GPDM (Fig.1(f)) for two different motion activities: salsa dance and exercises and quantitatively compare it against two state of art techniques, GPLVM [6] and GPDM [12].

For salsa dance, we roughly segment the sequence into 3 primitive motions: turn clockwise, turn anticlockwise and miscellaneous motion (e.g., move fwd/bwd, twist body). We use two synchronized camera views at front and side, and for each silhouette image, we take 10-dim PCA features and 10-dim moment features [10]. The test prediction

errors (MSE) and the selected frame results are shown in Fig.4 for the three techniques.

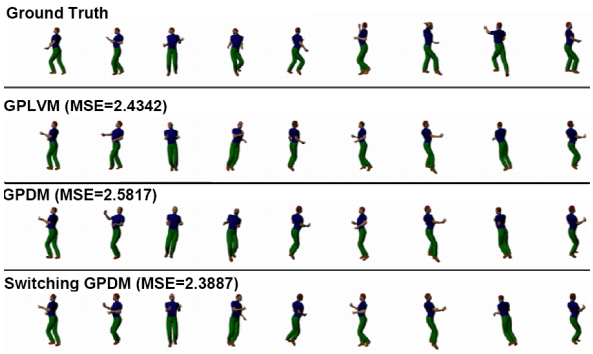


Figure 4. Salsa dance motion tracking results and tracking errors for three techniques.

The challenging exercise motion consists of 6 different primitive motions: jumping jack, turn-torso, turn-arms, hand-to-foot, knee-to-elbow, squat-up. We use the front-view silhouette images for measurement, where we extracted the 10-dim PCA features and the 10-dim moment features. The test prediction errors and some selected synthesized images are shown below in Fig.5.

Figure 5. Exercise motion tracking results and tracking errors for three techniques.

Both Figures 4 and 5 show that the proposed Switching GPDM can switches automatically among different dynamics, and outperforms GPLVM and GPDM visually and quantitatively. The performance improvement is especially significant for the more complex exercise motion sequence.

4.2. Facial Feature Tracking and Expression/Pose Recognition

We can easily apply the Switching GPDM to the simultaneous facial feature tracking and facial expression/pose recognition problem. Here we track 28 facial feature points (X) on the face (as shown in Fig.6), and the switching variables (S) now have explicit meaning of facial expression (neutral, surprise, happy, disgust and fear, as shown in Fig.7) and/or the out-of-plane facial poses.

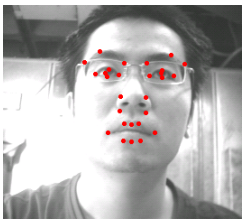


Figure 6. 28 facial feature points around eye, eyebrow, nose and mouth.

Figure 7. Five face expressions as the switching states.

Different from the silhouette measurement in human body motion tracking, we extract the Gabor features around the facial feature points as our measurement. Hence, we use the Switching GPDM in Fig.1(e) for facial feature tracking, and directly model the observation likelihood $p(\mathbf{v}_t|\mathbf{x}_t)$ based on an off-line trained Gabor template matching:

$$p(\mathbf{v}_t|\mathbf{x}_t) = p(\mathbf{v}_t|\mathbf{x}_t, I_t, \mathbf{T}) \approx e^{S_a(G(\mathbf{x}_t, I_t), \mathbf{T})} \quad (16)$$

where, $G(\mathbf{x}_t, I_t)$ is Gabor feature extracted from the current image I_t, $\mathbf{T}$ is the off-line trained Gabor template. $S_a(\bullet, \bullet) \in [-1, 1]$ is the similarity functions which compares two Gabor features. (More details about Gabor feature extraction and matching can be found in [14]).

Since the observation likelihood is directly defined, the learning process of Switching GPDM becomes independent of the image observations. Given the labeled facial feature point positions and expressions: $\{\mathbf{x}_t, s_t\}_{t=1}^{T}$, the model is learned by the method given in section 3.3.1.

With this model, facial feature tracking and facial expression/pose recognition are performed simultaneously in a unified framework. Such a framework represents a significant change from most current practices, which tend to treat the two problems of facial feature tracking and facial expressions separately and therefore ignore their interplays and inter-dependencies. Moreover, inference (sampling) in the low-dimensional latent space can improve the tracking accuracy with significantly reduced computational complexity.

In the experiments, we demonstrate the robustness of our model in terms of both facial feature tracking and expres-

sion/pose recognition performance on real videos under diverse scenarios including low frame rate and low quality videos, and face with out-of-plane pose changes.

4.2.1 Facial feature tracking + Expression Recognition

In this experiment, the model is trained on image sequences (each with length around 450) of one subject and then tests on the other sequences. 28 facial points and 5 expression expressions are manually labeled. We also down-sampled the sequence (sample every 5^{th} frame, the video frame rate is reduced from 30fps to 6fps), so as to simulate the low-frame rate videos.

The motivations for experimenting on low frame rate(LFR) videos include: (1) LFR tracking is more challenging for previous trackers which assume that $\mathbf{x}_t$ is quite close to $\mathbf{x}_{t-1}$ (2) Because the dynamic model of the switching state is trained by all the (s_t, s_{t-1}) pairs in the training data, reducing the lengths of equi-state segments might yield a better dynamic model. The tracking results are shown in Table 1:

Table 1. Tracking errors for different facial components for the proposed Switching GPDM (SGPDM) and the tracker in [11] based on Gabor feature and ASM under two different frame rates (without pose change)

Video Type	Model	Tracking Errors (per pixel)						
		L-EBR	L-EYE	R-EBR	R-EYE	NOSE	MOUTH	ALL
High Frame Rate	ASM	2.00	0.94	1.54	1.02	1.18	1.84	1.42
	SGPDM	2.88	1.38	2.51	1.36	2.62	3.71	2.50
Low Frame Rate	ASM	1.64	1.25	2.01	1.42	1.50	3.85	2.18
	SGPDM	2.14	1.26	2.31	1.25	2.63	3.94	2.43

Here, we compared with the state-of-the-art face tracker [11] based on the active shape model (ASM). We can see from the table, ASM-based tracker performs better than SGPDM on HFR (30fps) video. However, for LFR video, the tracking performances of SGPDM and ASM based tracker become comparable. Meanwhile, the SGPDM also yields the expression recognition result. The recognition error is 42% for HFR video and 22% for LFR video. This verify the assumption that LFR produces a better GP dynamic model.

4.2.2 Facial feature tracking+ Expression/Pose Recognition

Now, we take into account out-of-plane head poses. We roughly categorize pose into 3 groups: frontal, left, right, and deal with two expressions: neutral and happy. We form a switching state as a joint pose/expression, that is, we have 6 states (e.g., neutral-front). The tracking results are summarized in Table.2 and the expression/pose prediction errors are 33% for HFR and 29% for LFR.

Table 2. Tracking errors for different facial components for SGPDM and the tracker in [11] based on Gabor feature and ASM under two different frame rates (with pose change)

Video Type	Model	Tracking Errors (per pixel)						
		L-EBR	L-EYE	R-EBR	R-EYE	NOSE	MOUTH	ALL
High Frame Rate	ASM	4.76	2.48	4.19	4.36	2.31	2.08	3.10
	SGPDM	3.00	1.46	2.88	1.46	2.22	3.28	2.41
Low Frame Rate	ASM	4.84	3.95	3.61	4.28	3.81	3.51	3.92
	SGPDM	2.91	1.44	2.81	1.55	2.32	3.34	2.43

From Table.2, it is clear that the proposed method significantly outperforms the state-of-the-art face tracker for every facial component for both low and high frame rate videos, especially for the low frame rate video.

To further investigate the performance of our method, we did the experiment with low resolution images. We run the low-resolution video experiments by imputing every (5×5) site by the pixel value of their upper-left pixel. The results are summarized in Table 3.

Table 3. Average tracking errors and expression/pose recognition accuracies (in parenthesis) for different image conditions.

Expr + Pose	High Frame Rate		Low Frame Rate	
	High Res.	Low Res.	High Res.	Low Res.
ASM	3.10	4.46	3.92	4.47
SGPDM	2.41 (67.48%)	3.19 (42.09%)	2.43 (71.11%)	3.47 (38.89%)

Table 3 shows that for sequences with expression and pose change, the proposed method outperforms the ASM based tracker in every cases. In addition, we notice that while the expression/pose recognition of SGPDM is affected significantly by the low resolution image, the facial feature tracking errors are not affected as much.

5. Conclusion

In this paper, we introduce a dynamic model that can simultaneously address two problems in motion tracking, namely tracking complexity and tracking composite motions. Based on combining Gaussian Process Dynamic Models and switching variables, the proposed SGPDM model can efficiently track composite motions with diverse dynamics. The switching variables allows systematically capture the interplays and dependencies between high

level motion patterns and the dynamics of each motion pattern. And the Gaussian Process Dynamic Models reduce modeling complexity by modeling the dynamics in a low-dimensional latent space, improving both tracking accuracy and efficiency. The proposed SGPDM is also more suitable for the problem of simultaneous tracking and recognition. Experiments with human body motion tracking reveal the improved tracking performance of the proposed technique over the state-of-the-art methods. Experiments with facial feature tracking also indicate that SGPDM is suitable for the problem of simultaneous tracking and recognition.

References

[1] F. Dornaika and F. Davoine. Simultaneous facial action tracking and expression recognition in the presence of head motion. *IJCV*, 2008.

[2] K. Grochow, S. L.Martin, A. Hertzmann, and Z. Popvic. Style-based inverse kinematics. *SIGGRAPH*.

[3] M. Isard and A. Blake. A mixed-state condensation tracker with automatic model-switching. *ICCV*, 1998.

[4] N. D. Lawrence. Probabilistic non-linear principal component analysis with gaussian prcess latent variable models. *Journal of Machine Learning Research (JMLR)*, 2005.

[5] J. M.Wang, D. J. Fleet, and A. Hertzmann. A gaussian process dynamical models for human motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (2):283–298, 2008.

[6] R. Navaratnam, A. W.Fitzgibbon, and R. Cipolla. The joint mnifold model for semi-supervised multi-valued regression. *ICCV*, 2007.

[7] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *PAMI*, 2000.

[8] V. Pavlovic, J. M. Rehg, , T.-J. Cham, and K. P. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. *IEEE International Conference on Computer Vision*, 1999.

[9] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[10] T.-P. Tian, R. Li, and S. Sclaroff. Articulated pose estimation in a learned smooth space of feasible solutions.

[11] Y. Tong, Y. Wang, Z. Zhu, and Q. Ji. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recogn.*, 2007.

[12] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. *CVPR*, 2006.

[13] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. *ICCV*, 2005.

[14] L. Wiskott. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:775–779, 1997.

[15] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *CVIU*, 2003.