A Dynamic Bayesian Network Approach to Figure Tracking Using Learned Dynamic Models

Vladimir Pavlović, James M. Rehg, and Tat-Jen Cham
Compaq Computer Corporation
Cambridge Research Lab
Cambridge, MA 02139
{vladimir,rehg,tjc}@crl.dec.com

Kevin P. Murphy
Computer Science Division
U.C. Berkeley
Berkeley, CA 94720
murphyk@cs.berkeley.edu

Abstract

The human figure exhibits complex and rich dynamic behavior that is both nonlinear and time-varying. However, most work on tracking and synthesizing figure motion has employed either simple, generic dynamic models or highly specific hand-tailored ones. Recently, a broad class of learning and inference algorithms for time-series models have been successfully cast in the framework of dynamic Bayesian networks (DBNs). This paper describes a novel DBN-based switching linear dynamic system (SLDS) model and presents its application to figure motion analysis. A key feature of our approach is an approximate Viterbi inference technique for overcoming the intractability of exact inference in mixed-state DBNs. We present experimental results for learning figure dynamics from video data and show promising initial results for tracking, interpolation, synthesis, and classification using learned models.

1 Introduction

The human figure exhibits complex and rich dynamic behavior. Dynamics are essential to the analysis of human motion (e.g. gesture recognition) as well as to the synthesis of realistic figure motion in computer graphics. In visual tracking applications, dynamics can provide a powerful cue in the presence of occlusions and measurement noise.

Although the use of kinematic models in figure tracking is now commonplace, dynamic models have received relatively little attention. The kinematics of the figure specify its degrees of freedom (e.g. joint angles and torso pose) and define a state space. A dynamic model imposes additional structure on the state space by specifying which state trajectories are possible (or probable) and by specifying the speed at which a trajectory evolves.

One approach to dynamic modeling comes from the field of biomechanics. From this point of view, the dynamics of the figure are the result of its mass distribution, joint torques produced by the motor control system, and reaction forces resulting from contact with the environment (e.g. the floor). Research efforts in biomechanics, rehabilitation, and sports medicine have resulted in complex, specialized models of human motion. For example, entire books have been written on the subject of walking [10]. This approach has also been used successfully to produce computer graphics animations [9] and to track upper body motion in a user-interface setting [23].

The biomechanical approach has two drawbacks for visual tracking applications. First, the dynamics of the figure are quite complex, involving a large number of masses and applied torques, along with reaction forces which are difficult to measure using only visual data. In principle all of these factors must be modeled or estimated in order to produce physically-valid dynamics. Second, in some applications we may only be interested in a small set of motions, such as a vocabulary of gestures. In the biomechanical approach it may be difficult to reduce the complexity of the model to exploit this restricted focus.

This paper explores the alternative method of learning dynamic models from a training corpus of observed state space trajectories. In cases where sufficient training data is available, the learning approach promises flexibility and generality. A wide range of learning algorithms can be cast in the framework of Dynamic Bayesian Networks (DBNs). DBNs generalize two well-known signal modeling tools: Kalman filters [1] for continuous state linear dynamic systems (LDS) and Hidden Markov Models (HMMs) [20] for classification of discrete state sequences.

The DBN framework provides two distinct benefits: First, a broad variety of modeling schemes can be conceptualized in a single framework with an intuitively-appealing graphical notation (see Figure 1 for an example). Second, a broad corpus of exact and approximate statistical inference and learning techniques from the BN literature can be applied to dynamical systems. In particular, it has been shown that estimation in LDSs and inference in HMMs are special cases of inference in DBNs.

The focus of this paper is on a subclass of DBN models called Switching Linear Systems [2, 22, 14, 8, 19]. Intuitively, these models attempt to describe a complex nonlinear dynamic system with a succession of linear models that are indexed by a switching variable. While other approaches such as learning weighted combinations of linear models are possible, the switching approach has an appealing simplicity and is naturally suited to the case where the dynamics are time-varying.

This paper makes two contributions. First, we demonstrate the application of the SLDS framework to modeling figure dynamics. In particular, we demonstrate the learning of switching models of fronto-parallel walking and jogging motion from video data. We demonstrate the application of these learned models to segmentation and tracking tasks. Second, we derive a mixed-state version of the Viterbi approximation algorithm for inference in DBNs. Efficient approximation schemes are crucial for the practical application of these models. Our results demonstrate the promise of the SLDS approach to modeling visual dynamics.

2 Switching Linear System Model

Consider a complex physical system whose parameters evolve in time according to some known model. The system can be described using the following set of state-space equations:

$$\begin{aligned} x_{t+1} &= A(s_{t+1})x_t + v_{t+1}(s_{t+1}), \\ y_t &= Cx_t + w_t, \text{ and} \\ x_0 &= v_0(s_0) \end{aligned}$$

for the physical system, and

$$\begin{aligned} Pr(s_{t+1}|s_t) &= s'_{t+1}\Pi s_t, \text{ and} \\ Pr(s_0) &= \pi_0 \end{aligned}$$

for the switching model. The meaning of the variables is as follows: $x_t \in \Re^N$ denotes the hidden state of the LDS, and v_t is the state noise process. Similarly, $y_t \in \Re^M$ is the observed measurement and w_t is the measurement noise. Parameters A and C are the typical LDS parameters: the state transition matrix and the observation matrix, respectively. We assumed that the LDS models a Gauss-Markov process. Hence, the noise processes are independently distributed Gaussian:

$$\begin{aligned} v_t(s_t) &\sim \mathcal{N}(0, Q(s_t)),\ t > 0 \\ v_0(s_0) &\sim \mathcal{N}(x_0(s_t), Q_0(s_t)) \\ w_t &\sim \mathcal{N}(0, R). \end{aligned}$$

The switching model is assumed to be a discrete first order Markov process. State variables of this model are written as s_t. They belong to the set of S discrete symbols $\{e_0, \ldots, e_{S-1}\}$, where e_i is the unit vector of dimension S with a non-zero element in the i-th position. The switching model is defined with the state transition matrix Π whose elements are $\Pi(i,j) = Pr(s_{t+1} = e_i | s_t = e_j)$, and an initial state distribution π_0.

Coupling between the LDS and the switching process stems from the dependency of the LDS parameters A and Q on the switching process state s_t. Namely,

$$\begin{aligned} A(s_t = e_i) &= A_i \\ Q(s_t = e_i) &= Q_i \end{aligned}$$

In other words, switching state s_t determines which of S possible plant models is used at time t.

The complex state space representation is equivalently depicted by the DBN dependency graph in Figure 1 and can be written as the *joint distribution* P:

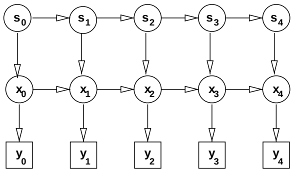


Figure 1: Bayesian network representation (dependency graph) of the SLDS. s denote instances of the discrete valued action states switching the physical system models with continuous valued states x and observations y.

$$\begin{aligned} P(\mathcal{Y}_T, \mathcal{X}_T, \mathcal{S}_T) = Pr(s_0) \textstyle\prod_{t=1}^{T-1} Pr(s_t|s_{t-1}) \\ Pr(x_0|s_0) \textstyle\prod_{t=1}^{T-1} Pr(x_t|x_{t-1}, s_t) \\ \textstyle\prod_{t=0}^{T-1} Pr(y_t|x_t). \end{aligned}$$

where $\mathcal{Y}_T$, $\mathcal{X}_T$, and $\mathcal{S}_T$ denote the sequences (of length T) of observations and hidden state variables. For instance, $\mathcal{Y}_T = \{y_0, \ldots, y_{T-1}\}$. We can write an equivalent representation of the physical system in the probability space assuming that the necessary conditional pdfs are defined. In fact they are. From the Gauss-Markov assumption on the LDS it follows:

$$\begin{aligned} x_{t+1}|x_t, s_{t+1} = e_i &\sim \mathcal{N}(A_i x_t, Q_i), \\ y_t|x_t &\sim \mathcal{N}(Cx_t, R), \\ x_0|s_0 = e_i &\sim \mathcal{N}(x_{0,i}, Q_{0,i}) \end{aligned}$$

Recalling the Markov switching model assumption, the joint pdf of the complex DBN of duration T (or, equivalently, its Hamiltonian[1]) can be written as in Equation 1.

[1] Hamiltonian $H(x)$ of a distribution $P(x)$ is defined as any positive function such that $P(x) = \frac{\exp(-H(x))}{\sum_\psi \exp(-H(\psi))}$.

$$H(\mathcal{X}_T, \mathcal{S}_T, \mathcal{Y}_T) = \frac{1}{2} \sum_{t=1}^{T-1} \sum_{i=0}^{N-1} \left[(x_t - A_i x_{t-1})' \, {Q_i}^{-1} \, (x_t - A_i x_{t-1}) + \log |Q_i| \right] s_t(i) +$$
$$\frac{1}{2} \sum_{i=0}^{N-1} \left[(x_{0,i})' \, {Q_{0,i}}^{-1} \, (x_{0,i}) + \log |Q_{0,i}| \right] s_0(i) + \tfrac{NT}{2} \log 2\pi +$$
$$\frac{1}{2} \sum_{t=0}^{T-1} (y_t - C x_t)' \, R^{-1} \, (y_t - C x_t) + \tfrac{T}{2} \log |R| + \tfrac{MT}{2} \log 2\pi$$
$$+ \sum_{t=1}^{T-1} s_t' (-\log \Pi) s_{t-1} + s_0' (-\log \pi_0). \tag{1}$$

2.1 Hidden State Inference

The goal of inference in complex DBNs is to estimate the posterior probability of the hidden states of the system (s_t and x_t) given some known sequence of observations $\mathcal{Y}_T$ and the known model parameters. Namely, we need to find the posterior

$$P(\mathcal{X}_T, \mathcal{S}_T | \mathcal{Y}_T) = Pr(\mathcal{X}_T, \mathcal{S}_T | \mathcal{Y}_T),$$

or its *sufficient statistics*. Given the form of P it is easy to show that these statistics are $\langle [x_t s_t] \rangle$, $\langle [x_t s_t][x_t s_t]' \rangle$, and $\langle [x_t s_t][x_{t-1} s_{t-1}]' \rangle$[2].

If there were no switching dynamics, the inference would be straightforward – we could infer $\mathcal{X}_T$ from $\mathcal{Y}_T$ using LDS inference (RTS smoothing [21]). However, the presence of switching dynamics embedded in matrix Π makes exact inference more complicated. To see that, assume that the initial distribution of x_0 at $t = 0$ is Gaussian, at $t = 1$ the pdf of the physical system state x_1 becomes a mixture of S Gaussian pdfs since we need to marginalize over S possible but unknown plant models. At time t we will have a mixture of S^t Gaussians, which is clearly intractable for even moderate sequence lengths. So, it is more plausible to look for an approximate, yet tractable, solution to the inference problem.

2.2 Approximate Inference Using Viterbi Approximation

The task of Viterbi approximation approach is to find the best sequence of switching states s_t and LDS states x_t that minimizes the Hamiltonian cost in Equation 1 for a given observation sequence $\mathcal{Y}_T$. It is well known how to apply Viterbi inference to discrete state hidden Markov models [20] and continuous state Gauss-Markov models [13]. Here we develop an algorithm for Viterbi inference in SLDSs (complex discrete/continuous DBNs.)

Define first the "best" *partial cost* up to time t of the measurement sequence $\mathcal{Y}_t$ when the switch is in state i at time t:

$$J_{t,i} = \min_{\mathcal{S}_{t-1}, \mathcal{X}_t} H\left(\{\mathcal{S}_{t-1}, s_t = e_i\}, \mathcal{X}_t, \mathcal{Y}_t \right) \tag{2}$$

Namely, this cost is the least cost over all possible sequences of switching states $\mathcal{S}_{t-1}$ and corresponding LDS states $\mathcal{X}_t$. This partial cost is essential in Viterbi-like total cost minimization. In order to calculate this cost we first define the following LDS state and variance terms:

$$\begin{aligned}
\hat{x}_{t|t,i} &\triangleq \langle x_t | \mathcal{Y}_t, s_t = e_i \rangle \\
\Sigma_{t|t,i} &\triangleq \left\langle (x_t - \hat{x}_{t|t,i})(x_t - \hat{x}_{t|t,i})' | \mathcal{Y}_t, s_t = e_i \right\rangle \\
\hat{x}_{t|t-1,i,j} &\triangleq \langle x_t | \mathcal{Y}_{t-1}, s_t = e_i, s_{t-1} = e_j \rangle \\
\Sigma_{t|t-1,i,j} &\triangleq \left\langle (x_t - \hat{x}_{t|t-1,i,j})(x_{t-1} - \hat{x}_{t|t-1,i,j})' | \right. \\
&\qquad \left. \mathcal{Y}_{t-1}, s_t = e_i, s_{t-1} = e_j \right\rangle \\
\hat{x}_{t|t,i,j} &\triangleq \langle x_t | \mathcal{Y}_t, s_t = e_i, s_{t-1} = e_j \rangle \\
\Sigma_{t|t,i,j} &\triangleq \left\langle (x_t - \hat{x}_{t|t,i})(x_t - \hat{x}_{t|t,i})' | \right. \\
&\qquad \left. \mathcal{Y}_t, s_t = e_i, s_{t-1} = e_j \right\rangle
\end{aligned}$$

$\hat{x}_{t|t,i}$ is the "best" filtered LDS state estimate at t when the switch is in state i at time t and a sequence of t measurements, $\mathcal{Y}_t$, has been processed. $\hat{x}_{t|t-1,i,j}$ and $\hat{x}_{t|t,i,j}$ are the one-step predicted LDS state and the "best" filtered state estimates at time t, respectively, given that the switch is in state i at time t and in state j at time $t-1$ and only $t-1$ measurements are known. Similar definitions are used for filtered and predicted state variance estimates, $\Sigma_{t|t,i}$ and $\Sigma_{t|t-1,i,j}$ respectively.

For a given switch state transition $j \to i$ it is now easy to establish relationship between the filtered and the predicted estimates. From the theory of Kalman estimation (see [1], for example) it follows that for transition $j \to i$ the following *time updates* hold:

$$\hat{x}_{t|t-1,i,j} = A_i \hat{x}_{t-1|t-1,j} \tag{3}$$
$$\Sigma_{t|t-1,i,j} = A_i \Sigma_{t-1|t-1,j} A_i' + Q_i. \tag{4}$$

Given a new observation y_t at time t each of these predicted estimates can now be filtered using Kalman *measurement update* framework. For instance, the state estimate measurement update equation yields

$$\hat{x}_{t|t,i,j} = \hat{x}_{t|t-1,i,j} + K_{i,j} \left(y_t - C \hat{x}_{t|t-1,i,j} \right). \tag{5}$$

[2] The operator $\langle \cdot \rangle$ denotes conditional expectation with respect to the posterior distribution, e.g. $\langle x_t \rangle = \sum_{\mathcal{S}} \int_{\mathcal{X}} x_t P(\mathcal{X}, \mathcal{S} | \mathcal{Y})$.

$$J_{t,t-1,i,j} = \frac{1}{2}\left(y_t - C\hat{x}_{t|t-1,i,j}\right)'\left(C\Sigma_{t|t-1,i,j}C' + R\right)^{-1}\left(y_t - C\hat{x}_{t|t-1,i,j}\right) + \frac{1}{2}\log\left|C\Sigma_{t|t-1,i,j}C' + R\right| - \log\Pi(i,j) \quad (6)$$

Appropriate equations can be obtained that link $\Sigma_{t|t-1,i,j}$ and $\Sigma_{t|t,i,j}$.

Each of these $j \to i$ transitions has a certain *innovation cost* $J_{t,t-1,i,j}$ associated with it, as defined in Equation 6. One portion of the innovation cost reflects the LDS state transition, as indicated by the innovation terms in Equation 6. The remaining cost is due to switching from state j to state i, $-\log\Pi(i,j)$.

Obviously, for every current switching state i there are S possible previous switching states where the system could have originated from. To minimize the overall cost at every time step t and for every switching state i one "best" previous state j is selected:

$$J_{t,i} = \min_j \left\{J_{t,t-1,i,j} + J_{t-1,j}\right\} \quad (7)$$

$$\psi_{t-1,i} = \arg\min_j \left\{J_{t,t-1,i,j} + J_{t-1,j}\right\}. \quad (8)$$

The index of this state is kept in the state transition record $\psi_{t-1,i}$. Consequently, we now obtain a set of S best filtered LDS states and variances at time t:

$$\begin{aligned} \hat{x}_{t|t,i} &= \hat{x}_{t|t,i,\psi_{t-1,i}} && (9) \\ \Sigma_{t|t,i} &= \Sigma_{t|t,i,\psi_{t-1,i}}. && (10) \end{aligned}$$

Once all T observations $\mathcal{Y}_{T-1}$ have been fused the best overall cost is obtained as

$$J^*_{T-1} = \min_i J_{T-1,i}. \quad (11)$$

To decode the "best" switching state sequence one uses the index of the best final state, $i^*_{T-1} = \arg\min_i J_{T-1,i}$, and then traces back through the state transition record $\psi_{t-1,i}$,

$$i^*_t = \psi_{t,i^*_{t+1}}. \quad (12)$$

Switching model's sufficient statistics are now simply $\langle s_t \rangle = e_{i^*_t}$ and $\langle s_t s'_{t-1} \rangle = e_{i^*_t} e'_{i^*_{t-1}}$. Given the "best" switching state sequence the sufficient LDS statistics can be easily obtained using the Rauch-Tung-Streiber smoothing [21]. For example,

$$\langle x_t, s_t(i) \rangle = \begin{cases} \hat{x}_{t|T-1,i^*_t} & i = i^*_t \\ 0 & \text{otherwise} \end{cases}$$

for $i = 0, \ldots, S-1$.

The Viterbi inference algorithm for complex DBNs can now be summarized as:

```
Initialize LDS state estimates $\hat{x}_{0|-1,i}$ and $\Sigma_{0|-1,i}$;
Initialize cost $J_{0,i}$.
for $t = 1 : T-1$
    for $i = 1 : S$
        for $j = 1 : S$
            Predict and filter LDS state estimates
              $\hat{x}_{t|t,i,j}$ and $\Sigma_{t|t,i,j}$
            Find innovation cost $J_{t|t-1,i,j}$
        end
        Find "best" partial cost $J_{t,i}$, state transition
          $\psi_{t-1,i}$, and LDS state estimates
          $\hat{x}_{t|t,i}$ and $\Sigma_{t|t,i}$
    end
end
Find "best" final switching state $i^*_{T-1}$.
Backtrack to find "best" switching state sequence $i^*_t$.
Find DBN's sufficient statistics.
```

2.3 Maximum Likelihood Learning of Complex DBNs

Learning in complex DBNs can be formulated as the problem of ML learning in general Bayesian networks. Hence, a generalized EM algorithm [17] can be used to find optimal values of DBN parameters $\{A, C, Q, R, \Pi, \pi_0\}$. The expectation (E) step of EM is the inference itself—we dealt with this task in the previous section.

Given the sufficient statistics from the inference phase, it is easy to obtain *parameter update equations* in the maximization (M) step. For instance, updated values of the state transition parameters are easily shown to be

$$\begin{aligned} \hat{A}_i &= \left(\sum_{t=1}^{T-1} \langle x_t x'_{t-1} s_t(i) \rangle\right) \\ &\quad \left(\sum_{t=1}^{T-1} \langle x_{t-1} x'_{t-1} s_t(i) \rangle\right)^{-1} \\ \hat{\Pi} &= \left(\sum_{t=1}^{T-1} \langle s_t s'_{t-1} \rangle\right) \operatorname{diag}\left(\sum_{t=1}^{T-1} \langle s_t \rangle\right)^{-1}. \end{aligned}$$

All the variable statistics are evaluated before updating any parameters. Notice that the above equations represent a generalization of the parameter update equations of classical (non-switching) LDS models [7].

3 Previous Work

Most previous figure trackers which have used a dynamic model employed a simple smoothness prior such as a constant velocity Kalman filter [12]. One exception is [23], in which an input estimation approach is used to estimate the joint torques in a 3-D dynamic model of the upper body. In the following section we demonstrate the superiority of our learned models over simple smoothness priors. We believe these methods provide a useful alternative to detailed biomechanical modeling.

SLDS models and their equivalents have been studied in statistics, time-series modeling, and target tracking since early 1970's. Bar-Shalom [2] and Kim [14] have developed a number of approximate pseudo-Bayesian inference techniques based on mixture component truncation or collapsing is SLDSs. They did not address the issue of learning system parameters. Shumway and Stoffer [22] presented a systematic view of inference and learning in SLDS while assuming known prior switching state distributions at each time instance, $Pr(s_t) = \pi_t(i)$ and no temporal dependency between switching states. Krishnamurthy and Evans [15] imposed Markov dynamics on the switching model. However, they assumed that noisy measurements of the switching states are available.

Ghahramani [8] introduced a DBN-framework for learning and approximate inference in one class of SLDS models. His underlying model differs from ours in assuming the presence of S independent, white noise-driven LDSs whose measurements are selected by the Markov switching process. An alternative input-switching LDS model was proposed by Pavlovic et al. [19] and utilized for mouse motion classification. A switching model framework for particle filters is described in [11] and applied to dynamics learning in [3]. Manifold learning [5] is another approach to constraining the set of allowable trajectories within a high dimensional state space. An HMM-based approach is described in [4].

4 Experiments

We applied our DBN-based SLDS framework to the analysis of two categories of fronto-parallel motion: walking and jogging. Fronto-parallel motions exhibit interesting dynamics and are free from the difficulties of 3-D reconstruction. Experiments can be conducted easily using a single video source, while self-occlusions and cluttered backgrounds make the tracking problem non-trivial.

We adopted the 2-D Scaled Prismatic Model proposed by Morris and Rehg [16] to describe the kinematics of the figure. The kinematic model lies in the image plane, with each link having one degree of freedom (DOF) in rotation and another DOF in length. A chain of SPM transforms can model the image displacement and foreshortening effects produced by 3-D rigid links. The appearance of each link in the image is described by a template of pixels which is manually initialized and deformed by the link's DOF's.

In our figure tracking experiments we analyzed the motion of the legs, torso, and head, and ignoring the arms. Our kinematic model had eight DOF's, corresponding to rotations at the knees, hip, and neck. A sample configuration of our figure model is shown in Figure 4.2.

4.1 Classification

The first task we addressed was learning an SLDS model for walking and running. Our training set consisted of 18 sequences of six individuals jogging (two examples of three people) and walking at a moderate pace (two examples of six people.) Each sequence was approximately 50 frames duration. The training data consisted of the joint angle states of the SPM in each image frame, which was obtained manually.

Each of the two motion types were each modeled as multi–state[3] SLDSs and then combined into a single complex SLDS. Measurement matrix in all cases was assumed to be identity, $C = I$. Initial state segmentation within each motion type was obtained using unsupervised clustering in a state space of some simple dynamics model (e.g. constant velocity model.) Parameters of the model $(A, Q, R, x_0, \Pi, \pi_0)$ were then reestimated using the EM-learning framework with approximate Viterbi inference. This yielded refined segmentation of switching states within each of the models. An example of the learned switching state sequence within a single "jog" training example is shown in Figure 2.

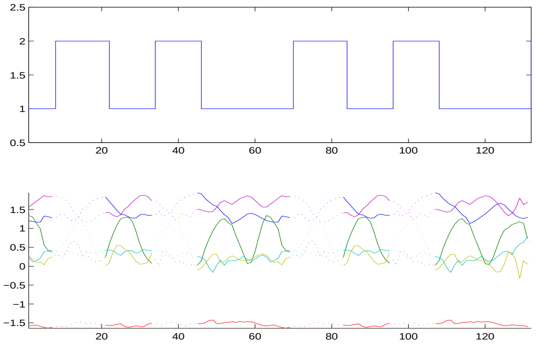


Figure 2: Segmentation of two-state SLDS model states within single "jog" motion sequence.

To test the classification ability of our learned model we next considered segmentation of sequences of *complex* motion, i.e., motion consisting of alternations of "jog" and "walk."[4] Identification of different motion "regimes" was

[3] We explored SLDS models with two to six states.

[4] Test sequences were constructed by concatenating in random order randomly selected and noise corrupted training sequences. Transitions between sequences were smoothed using B-spline smoothing.

conducted using the approximate Viterbi inference. Estimates of "best" switching states $\langle s_t \rangle$ indicated which of the two models can be considered to be driving the corresponding motion segment. One example of this segmentation is depicted in Figure 3. Classification experiments on a set of 20 test sequences gave an error rate[5] of 2.9% over a total of 8312 classified data points.

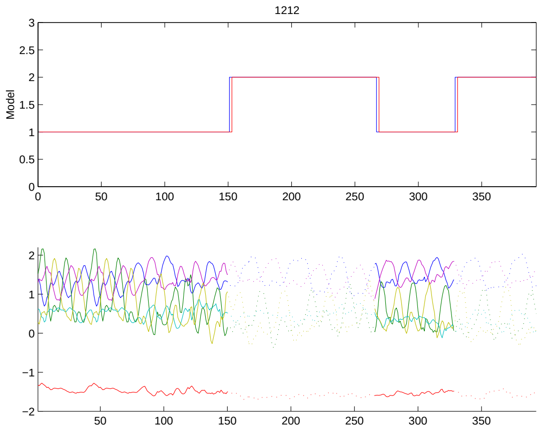


Figure 3: Segmentation of mixed walking/running sequence. Top graph shows correct segmentation (dotted red line) and estimated segmentation (solid blue line). Bottom graph depicts the segmentation of the estimated LDS states.

4.2 Tracking

A second experiment explored the utility of the SLDS model in improving tracking of the human figure from video. The difficulty in this case is that feature (joint angle) measurements are not readily available from a sequence of image intensities. Hence, we use the SLDS as a multi-hypothesis predictor that initializes multiple local template searches in the image space. Instead of choosing S^2 multiple hypotheses $\hat{x}_{t|t-1,i,j}$ at each time step as indicated in Equation 3 we pick the best S hypothesis with the smallest switching cost, i.e., $\hat{x}_{t|t-1,i,i_t^*}$ where $i_t^* = \arg\min_j \{-\log \Pi(i,j) + J_{t-1,j}\}$.

Given the predicted means for the figure locations, state-space observations are obtained by local image registration, or hill-climbing. This identifies the state-space modes in the likelihood function given by the template model. A larger set of measurements could be explored through sampling, as described in [6]. Given these observations of figure state, the regular SLDS filtering yields SLDS state priors.

Figure 4 shows stills from a representative example of SLDS tracking of walking motion. In this experiment, simple template features were used to model the appearance of the figure. Each link in the model has an associated template, which is initialized manually in the first frame and applied throughout the sequence. Template features are not robust to appearance changes such as lighting effects or the wrinkling of cloth. As a result, a template-based tracker can benefit substantially from an accurate dynamical model.

A constant velocity predictor does poorly in this case, leading to tracking failure by frame seven (shown in Figure 4.b). The learned SLDS model gives improved predictions leading to more robust tracking.

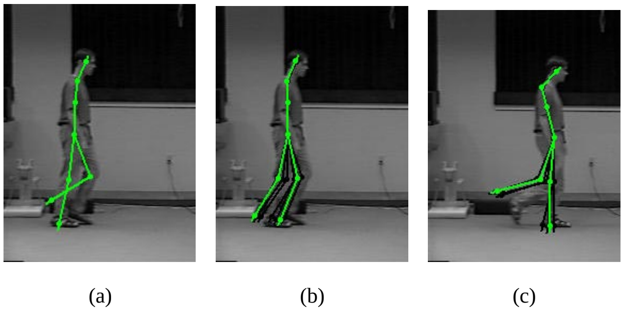

(a) (b) (c)

Figure 4: (a) Tracker (in white) using constant velocity predictor drifts off track by frame 7. (b) SLDS-based tracker is on track at frame 7. Model (switching state) 3 has the highest likelihood. Black lines show prior mean and observation. (c) SLDS tracker at frame 20.

4.3 Synthesis and Interpolation

In Section 2 we introduced SLDS as a *generative* model. Nonetheless, SLDS is most commonly employed as a classifier (e.g. Section 4.1.) To test the power of the learned SLDS framework we examined its use in synthesizing realistic–looking motion sequences and interpolating motion between missing frames.

In the first set of experiments the learned walk/jog SLDS was used to generate a "synthetic walk." Two stick figure motion sequences of the noise driven model are shown in Figure 5. Depending on the amount of noise used to drive the model the stick figure exhibits more or less "natural"–looking walk. Departure from the realistic walk becomes more evident as the simulation time progresses. This behavior is not unexpected as the SLDS in fact learns locally consistent motion patterns.

Another realistic situation may call for filling-in a small number of missing frames from a large motion sequence. SLDS can then be utilized as an interpolation function. In a set of experiments we employed the learned walk/jog model to interpolate a walk motion over two sequences with missing frames (see Figure 6.) The visual quality of

[5] Classification error was defined as the difference between inferred segmentation and true segmentation accumulated over all sequences, $e = \sum_{t=0}^{T-1} |\langle s_t \rangle - s_{\text{true},t}|$.

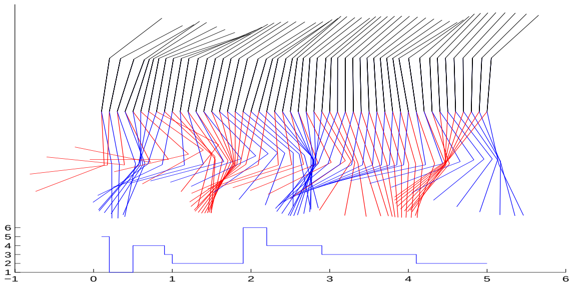


Figure 5: Synthesized walk motion over 50 frames using SLDS as a generative model. States of the synthesized motion are shown on the bottom. Right leg is in blue color.

the interpolation and the motion synthesized from it was high (left column in Figure 6.) As expected, the sparseness of the measurement set had definite bearing on this quality.

5 Conclusions

We have introduced a new approach to dynamics learning based on switching linear models. We have proposed a Viterbi approximation technique which overcomes the exponential complexity of exact inference. One open issue with this approach is the lack of an exact bound on the approximation error. This is a problem in general with greedy Viterbi-style approximations, as well as with Markov chain Monte Carlo methods [18]. One possibility alternative are the variational inference techniques used in [8, 19], which do have well-defined error bounds.

Our preliminary experiments have demonstrated promising results in classification of human motion, improved visual tracking performance, and motion synthesis and interpolation using our SLDS framework. We demonstrated accurate discrimination between walking and jogging motions. We showed that SLDS models provide more robust tracking performance than simple constant velocity predictors. The fact that these models can be learned from data may be an important advantage in figure tracking, where accurate physics-based dynamical models may be prohibitively complex.

We are currently building a more comprehensive collection of frontoparallel human motion. We plan to build SLDS models for wide variety of motions and performers and evaluate their performance.

References

[1] B. D. O. Anderson and J. B. Moore, *Optimal filtering*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1979.

[2] Y. Bar-Shalom and X.-R. Li, *Estimation and tracking: principles, techniques, and software*. Storrs, CT: YBS, 1998.

[3] A. Blake, B. North, and M. Isard, "Learning multi-class dynamics," in *NIPS '98*, 1998.

[4] M. Brand, "Pattern discovery via entropy minimization," Tech. Rep. TR98-21, Mitsubishi Electric Research Lab, 1998. Available at *http://www.merl.com*.

[5] C. Bregler and S. M. Omohundro, "Nonlinear manifold learning for visual speech recognition," in *ICCV*, (Cambridge, MA), pp. 494–499, June 1995.

[6] T.-J. Cham and J. M. Rehg, "A multiple hypothesis approach to figure tracking," in *CVPR*, pp. 239–245, 1999.

[7] Z. Ghahramani, "Learning dynamic Bayesian networks," in *Adaptive processing of temporal information* (C. L. Giles and M. Gori, eds.), Lecture notes in artificial intelligence, Springer-Verlag, 1997.

[8] Z. Ghahramani and G. E. Hinton, "Switching state-space models." submitted for publication, 1998.

[9] J. K. Hodgins, W. L. Wooten, D. C. Brogan, and J. F. O'Brien, "Animating human athletics," in *Computer Graphics (Proc. SIGGRAPH '95)*, pp. 71–78, 1995.

[10] V. T. Inman, H. J. Ralston, and F. Todd, *Human Walking*. Williams and Wilkins, 1981.

[11] M. Isard and A. Blake, "A mixed-state CONDENSATION tracker with automatic model-switching," in *ICCV*, (Bombay, India), pp. 107–112, 1998.

[12] I. A. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection," in *CVPR*, (San Fransisco, CA), pp. 81–87, June 18-20 1996.

[13] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction," *Journal of Basic Engineering (ASME)*, vol. D, no. 83, pp. 95–108, 1961.

[14] C.-J. Kim, "Dynamic linear models with markov-switching," *Journal of Econometrics*, vol. 60, pp. 1–22, 1994.

[15] V. Krishnamurthy and J. Evans, "Finite-dimensional filters for passive tracking of markov jump linear systems," *Automatica*, vol. 34, no. 6, pp. 765–770, 1998.

[16] D. D. Morris and J. M. Rehg, "Singularity analysis for articulated object tracking," in *CVPR*, (Santa Barbara, CA), pp. 289–296, June 23–25 1998.

[17] R. M. Neal and G. E. Hinton, "A new view of the EM algorithm that justifies incremental and other variants," in *Learning in graphical models* (M. Jordan, ed.), pp. 355–368, Kluwer Academic Publishers, 1998.

[18] R. M. Neal, "Connectionist learning of belief networks," *Artificial Intelligence*, pp. 71–113, 1992.

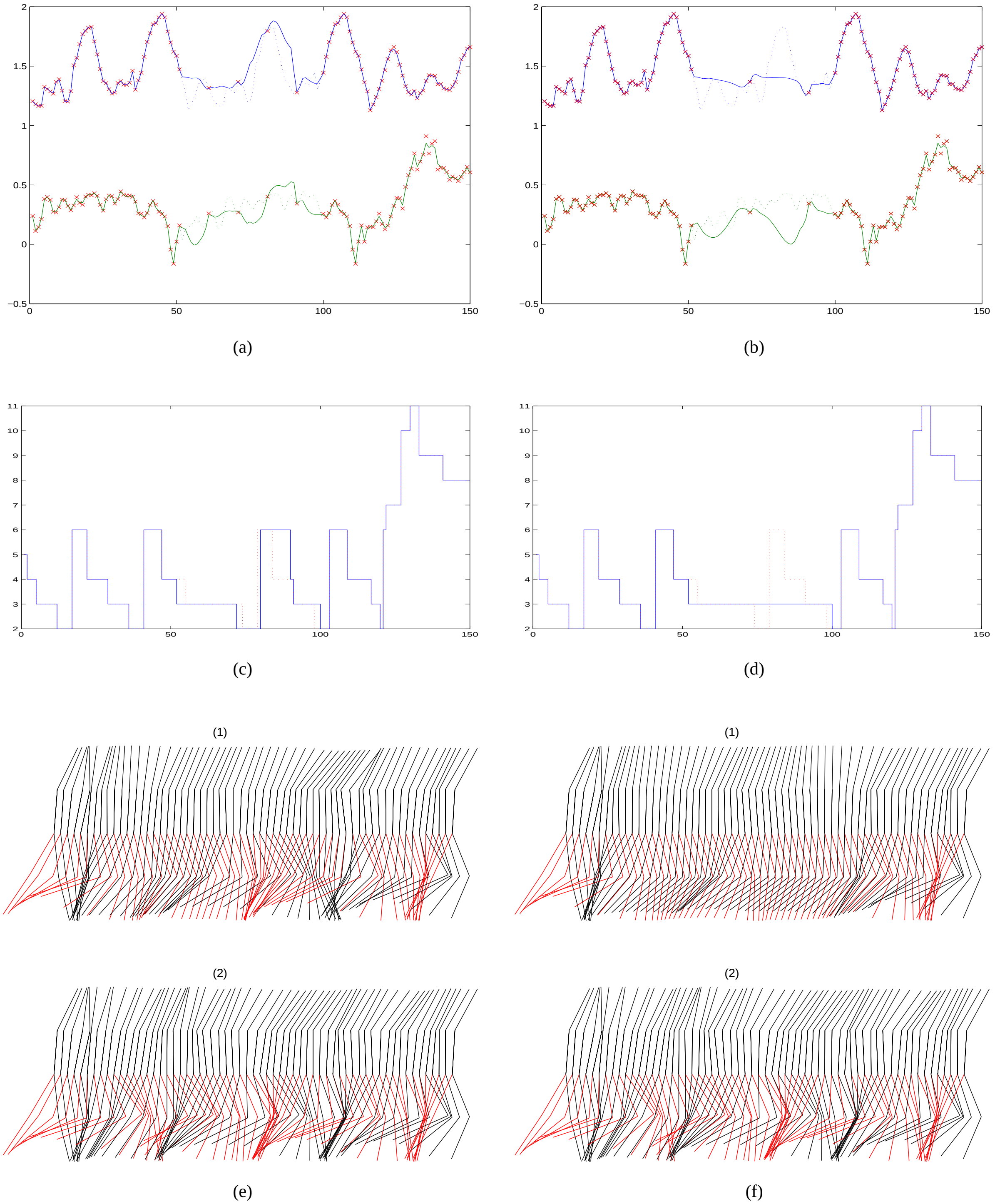
[19] V. Pavlovic, B. Frey, and T. S. Huang, "Time-series classification using mixed-state dynamic Bayesian networks," in *CVPR*, pp. 609–615, June 1999.

[20] L. R. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey, USA: Prentice Hall, 1993.

[21] H. E. Rauch, "Solutions to the linear smoothing problem," *IEEE Trans. Automatic Control*, vol. AC-8, pp. 371–372, October 1963.

[22] R. H. Shumway and D. S. Stoffer, "Dynamic linear models with switching," *Journal of the American Statistical Association*, vol. 86, pp. 763–769, September 1991.

[23] C. R. Wren and A. P. Pentland, "Dynamic models of human motion," in *Proc. 3rd International Conference on Automatic Face and Gesture Recognition*, (Nara, Japan), pp. 22–27, 1998.



(a) (b)

(c) (d)

(e) (f)

Figure 6: SLDS as an interpolation function. Two motion sequences (left and right column) with missing measurements between frames 50 and 100 were interpolated using an SLDS model. Symbols red 'x' in top two figures (a & b) indicate known measurement points. Solid lines show interpolated joint angle values. Dotted lines indicate ground truth (smoothing with no missing measurements.) Figures (c) and (d) depict corresponding SLDS states (blue for interpolated and red for ground truth.) Stick figure motion generated from interpolated data is shown in figures (e) and (f). Graphs (e.2) and (f.2) show true figure motion .