MDPI

Article

A Bayesian Dynamical Approach for Human Action Recognition

Amirreza Farnoosh, Zhouping Wang, Shaotong Zhu and Sarah Ostadabbas *

Augmented Cognition Lab (ACLab), Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA; afarnoosh@ece.neu.edu (A.F.); zpwang@ece.neu.edu (Z.W.); shawnzhu@ece.neu.edu (S.Z.)
* Correspondence: ostadabbas@ece.neu.edu

Abstract: We introduce a generative Bayesian switching dynamical model for action recognition in 3D skeletal data. Our model encodes highly correlated skeletal data into a few sets of low-dimensional switching temporal processes and from there decodes to the motion data and their associated action labels. We parameterize these temporal processes with regard to a switching deep autoregressive prior to accommodate both multimodal and higher-order nonlinear inter-dependencies. This results in a dynamical deep generative latent model that parses meaningful intrinsic states in skeletal dynamics and enables action recognition. These sequences of states provide visual and quantitative interpretations about motion primitives that gave rise to each action class, which have not been explored previously. In contrast to previous works, which often overlook temporal dynamics, our method explicitly model temporal transitions and is generative. Our experiments on two large-scale 3D skeletal datasets substantiate the superior performance of our model in comparison with the state-of-the-art methods. Specifically, our method achieved 6.3% higher action classification accuracy (by incorporating a dynamical generative framework), and 3.5% better predictive error (by employing a nonlinear second-order dynamical transition model) when compared with the best-performing competitors.

1. Introduction

Analyzing 3D motion capture datasets, illustrating dynamical motions of a subject and inferring their actions is the key processing step in many applications, including highlighting movement patterns of an athlete to optimize their performance, probing behavior of an endangered animal, and monitoring mobility of a patient in a rehabilitation study, to name a few [1]. In all these applications, body pose data contained in the motion capture sequence describes temporal evolution of specific phenomena or action and is switching between potentially limited number of states each representing a specific regime.
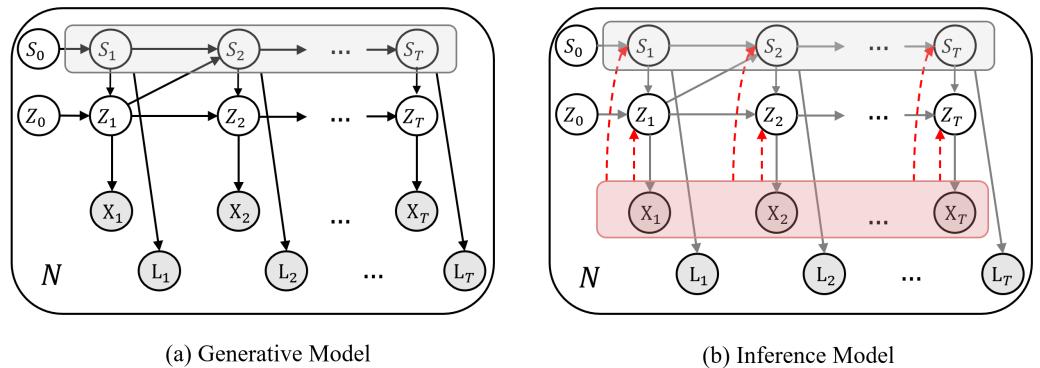
The efforts in quantifying complex kinematics of biological mechanisms in a lower dimensional subspace have led to the successful design of bio-inspired robots that can mimic their biological counterparts to a great extent [2–4]. For instance, it is shown that more than 80% of the variance of static grasping data in humans could be described by the first two postural synergies [5]. Skeleton-based human action recognition, on the other hand, has been extensively studied in the context of deep learning using convolutional neural networks [6,7], recurrent neural networks [8,9], and graph convolutional networks [10–12].

However, there are still some limitations with the existing body of works. (I) A vast majority of previous state-of-the art models perform action classification on motion segments that only include a single action. In other words, they predict a single action label for the entire motion segment. This substantially decreases their applicability in practice for real-world applications because in most cases (if not all) we are not given
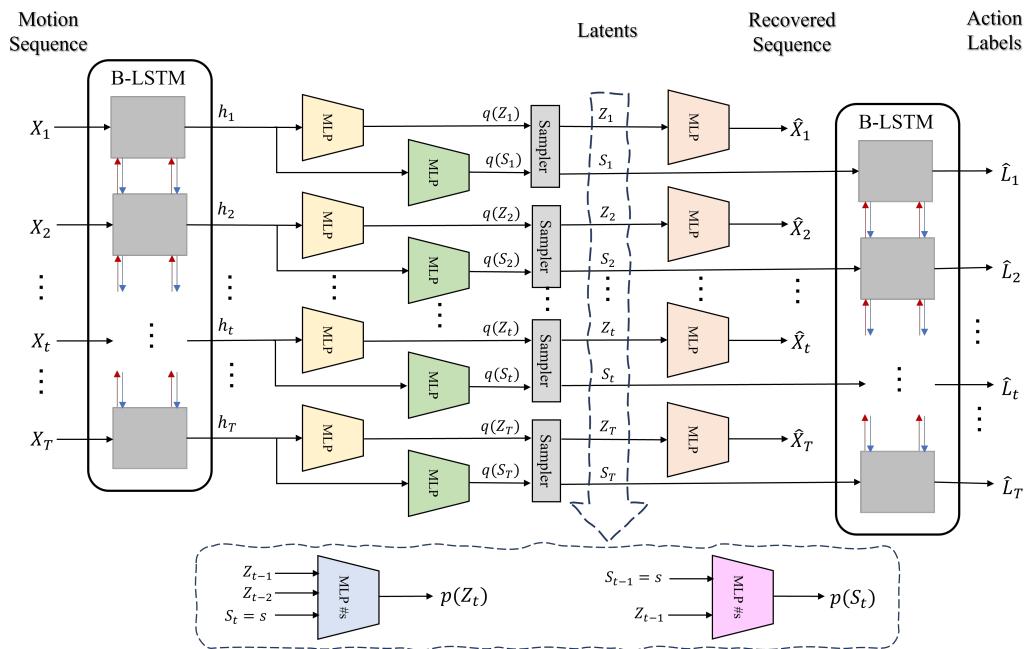
single-action segments and the action recognition model has to determine action segments itself. A practical algorithm should be able to provide per-time point action labels and update these labels accordingly in real-time. (II) Although some previous works capture temporal correlations in motion data (e.g., through temporal smoothness regularization), there are very few methods that explicitly model temporal dynamics/transitions. For the task of human action recognition, most of the existing models are deterministic, specifically designed for action recognition, and inherently do not model temporal dynamics or learn data distribution, and therefore are not generative. (III) Most of the previous methods cannot handle missing entries in motion data (which may happen frequently) because they are not predictive. Their common practice is to simply set missing entries to zero or remove their corresponding time points entirely. (IV) Most of the previous works do not provide any interpretations or intuitions for their predictions of action labels in terms of intrinsic dynamical processes in motion data or motion primitives.

The strong spatio-temporal correlation among joints of a human's skeleton captured by 3D motion capture data as well as its clear sparseness motivate the utilization of dynamical probabilistic models that can learn underlying interpretable states from data and extract their low-dimensional motion patterns. To this end, we propose a Bayesian state switching model for dynamical action recognition that is both generative and predictive. Specifically, we employ a low-dimensional deep generative latent model to encode highly correlated skeleton data into a few set of switching autoregressive temporal processes. This model then decodes from these low-dimensional representations to the skeletal data and their associated action labels. This results in a flexible model that accommodates multimodal and higher-order nonlinear inter-dependencies, parses meaningful temporal modes in 3D pose data, and enables action recognition. Specifically, we make the following contributions:

- In contrast to previous works which merely classify motion sequences into action labels and are not generative, our method, sketched in Figures 1 and 2, in addition to action recognition, (I) segments motion data from a dynamical perspective by explicitly modeling temporal dynamics in a Bayesian approach, hence, (II) it allows dynamical prediction of skeletal motion from low-level representations. Specifically, (III) it welcomes multimodal, higher-order, and nonlinear temporal relations in motion data by employing a deep switching auto-regressive latent model.

- Our method can easily fill in missing entries in motion data due to its predictive nature which models transitions between time points and between action labels.

- Our method provides action labels per time point and can handle varying-length sequences.

- Our method uses a nonlinear second-order dynamical model to better capture skeletal dynamics, because first-order models are less effective in modeling acceleration (i.e., second-derivative of location) in motion data.

- The sequence of discrete latents in our method enables multi-modal dynamical modeling and at the same time provides visual/qualitative and quantitative interpretations about motion primitives that gave rise to each action class, which were not possible with previous methods.

- Our method is probabilistic and provides confidence intervals for its estimations and predictions.

(a) Generative Model

(b) Inference Model

Figure 1. (**a**) Graphical representation of our generative model given N motion datasets X_n and their action labels L_n. The low-dimensional continuous latents $Z_n = \{Z_{n,t}\}_{t=1}^T$ are generated with regards to a nonlinear autoregressive prior switched by their associated discrete states $\mathcal{S}_n = \{\mathcal{S}_{n,t}\}_{t=1}^T$. The discrete states themselves are determined according to a Markovian prior conditioned on their preceding continuous latents, i.e., Z_{t-1}. (**b**) In the inference model, discrete and continuous latents are estimated from motion sequences X_n in an amortized fashion (denoted by dashed red arrows).


Figure 2. Visual framework of our proposed method. Our model encodes an input motion sequence $\{X_1, \cdots, X_T\}$ into a sequence of hidden features $\{h_1, \cdots, h_T\}$ using a bidirectional LSTM. The resulting hidden features are fed to two separate MLPs for estimating variational distribution parameters of discrete latents $\{\mathcal{S}_1, \cdots, \mathcal{S}_T\}$ and continuous latents, $\{Z_1, \cdots, Z_T\}$. These posterior distributions are then sampled to obtain their latent values. We decode to the input motion sequence $\{\hat{X}_1, \cdots, \hat{X}_t\}$ by feeding continuous latents Z to an MLP. We also decode to the associated action labels $\{\hat{L}_1, \cdots, \hat{L}_T\}$ by feeding probability vectors of the discrete latents $\mathcal{S}$ to a bidirectional LSTM. We estimate the priors for the discrete latents $p(\mathcal{S}_t)$ and continuous latents $p(Z_t)$ from the values of sampled latents using two separate MLPs.

We demonstrate superior performance of our model on two large-scale 3D skeletal motion datasets in terms of action classification and dynamical prediction accuracy. The source code is available at github.com/ostadabbas/ActionDSARF (accessed on 15 August 2021).

2. Related Works

Here, we give a short overview of the recent advancements in dynamical system modeling as well as human action recognition using skeletal data, two fundamental aspects of our proposed approach.

2.1. Dynamical Systems Modeling

Switching linear dynamical systems (SLDS) have long been investigated in the literature [13–18]. These models decompose time series data into series of simpler, repeated dynamical modes represented by discrete and continuous latent states. In SLDS framework, the transitions between discrete states are independent of their associated continuous values. This problem is addressed in recurrent switching linear dynamical system (rSLDS) [19–21], by allowing the discrete state transition probabilities to depend on their preceding continuous states. The rSLDS dynamical capacity is however limited as it assumes first-order linear dynamics. A recent work [22], extends these models by adopting nonlinear and higher-order multimodal dependencies through a deep switching autoregressive framework. We build our model on top of this framework and customize that for dynamical action recognition. This makes our model flexible for complex auto-regressive relations in motion sequences.

From another perspective, dynamical matrix factorization is used in [23–27] for modeling linear dynamics in their low-dimensional temporal factors. Several studies have also employed neural networks for non-linear state-space modeling [28–33], which are restricted to first-order Markovian dependencies, and for time series prediction [34–39], most of which are non-probabilistic.

2.2. Action Recognition

Skeleton-based human action recognition has drawn much attention for its efficiency and robustness compared to the image-based action recognition methods. Earlier works [40,41] treat joint coordinates as independent features and model the relationship between them through hand-crafted feature vectors. After the introduction of deep learning, both convolutional and recurrent networks are used to extract information from skeleton data after transforming them into pseudo images or a sequence of vectors. In order to accomplish this transformation both spatially and temporally, Ref. [6] proposed their joint trajectory maps to represent spatial configuration and joint dynamics using color-coded texture images. Authors in [7] proposed the shape-motion representation from geometric algebra, which addressed the importance of both joints and bones and fully utilized the information provided by skeleton sequence. Recurrent neural network (RNN) structure can effectively process sequential data, while it has more difficulties for spatial information modeling. A novel two-stream RNN architecture to model both temporal dynamics and spatial configurations for skeleton data is proposed in [8]. Human actions can also be interpreted based on the interactions between body parts. In [9], authors proposed part-aware long short-term memory (P-LSTM), which divides the entire body dynamics based on different body parts and learn the final classifier over their concatenation. Common temporal patterns of the parts are learnt independently and then combined in the global level representation for action recognition. Recent approaches model the joint movements of an action with skeleton spatial-temporal graphs. Authors in [10] proposed to model the skeleton data with the graph convolutional networks (GCNs), in which the spatial graph convolution and the temporal graph convolution are both utilized. Spatial edges correspond to the connection of joints at each frame and the temporal edges connect the same joints across frames. In this configuration, the spatio-temporal information can be extracted based on the multi-layer graph convolution. To extract discriminative spatio-temporal features more effectively, [11] proposed a novel graph-based LSTM network to capture discriminative features by exploring the co-occurrence relationship between spatial and temporal spaces. Hierarchical architectures are also used to increase the temporal receptive fields. In [12], the skeleton data are represented as a directed acyclic graph

(DAG) based on the kinematic dependency between the joints and bones in the natural human body to combine these two types of data for a better usage. In order to unbiasedly model long-range joint relationship under multi-scale operators and get unobstructed cross space-time information flow for capturing complex spatial-temporal dependencies, Ref. [42] proposed multi-scale aggregation scheme disentangling the importance of nodes in different neighborhoods for effective long-range modeling.

More recently, Ref. [43] proposed a bi-directional long short-term memory (BiLSTM) based attention mechanism with a dilated convolutional neural network (DCNN) for human action recognition in videos. They used DCNN layers equipped with residual blocks to extract discriminative features from video frames. These features are then fed into a BiLSTM network to learn temporal dependencies, which is followed by an attention mechanism to perform action classification. Also, for the task of speech emotion recognition, Ref. [44] proposed a two-stream deep convolutional neural network with an iterative neighborhood component analysis (INCA) to learn refined spatiotemporal features. They used two distinct convolutional neural networks, denoted as two streams, to extract spatial and spectral features separately. These features are then combined and fed to an INCA for further refinement. Finally, the jointly refined features are passed from a fully connected network with a softmax classifier to predict emotion categories.

3. Problem Formulation

We consider a set of N motion datasets $\{X_1, \ldots, X_N\}$ and their associated action labels $\{L_1, \ldots, L_N\}$, where each $X_n \in \mathbb{R}^{T \times (D \times 3)}$ contains 3D coordinates of D skeleton joints over T time points and $L_n \in \{1, \cdots, A\}^T$ denotes its per-time point action labels from a set of A action classes. We propose a low-dimensional deep latent model that learns the generative distribution of these data and infers their latent representations as well as actions labels for an unseen sequence, which we will explain below as "Generative" and "Inference" models, respectively.

3.1. Generative Model

We assume that each data pair $\{X_n, L_n\}$ is generated according to a set of *discrete* latent states $\mathcal{S}_n = \{\mathcal{S}_{n,t}\}_{t=1}^T$ representing motion primitives and their corresponding low-dimensional *continuous* temporal latent variables $Z_n = \{Z_{n,t} \in \mathbb{R}^K\}_{t=1}^T$:

$$
\begin{aligned}
X_n &\sim p_\theta(X_n \mid Z_n), \\
L_n &\sim p_\theta(L_n \mid \mathcal{S}_n), \\
Z_n &\sim p_\theta(Z_n \mid \mathcal{S}_n), \\
\mathcal{S}_n &\sim p_\theta(\mathcal{S}),
\end{aligned} \tag{1}
$$

where $p_\theta(X_n|Z_n)$ and $p_\theta(L_n|\mathcal{S}_n)$ are emission distributions that define conditional probabilities of observation space for motion data X_n and their action labels L_n with respect to local continuous latents Z_n and discrete latents $\mathcal{S}_n$, respectively. $p_\theta(Z_n|\mathcal{S}_n)$ is a switching dynamical autoregressive prior over Z_n and $p_\theta(\mathcal{S})$ is a generative Markovian prior over $\mathcal{S}_n$. We parameterize all distributions with neural networks and collectively denote their generative model parameters by θ. The graphical representation for our proposed generative model is depicted in Figure 1a.

3.1.1. Discrete Markovian Prior $p_\theta(\mathcal{S})$

We construct a predictive dynamical framework to capture temporal coherence in our sequential data. We assume that this dynamical generative process resides at a specific state at each time point (out of S possible states) which is determined according to a Markovian prior conditioned on its preceding discrete latent $\mathcal{S}_{t-1}$ and continuous latent Z_{t-1}. As such, the discrete latent states $\mathcal{S}_n = \{\mathcal{S}_{n,t}\}_{t=1}^T$ are structured in a Markov chain as follows:

$$
p_\theta(\mathcal{S}_t | \mathcal{S}_{t-1} = s, Z_{t-1}) = \text{Cat}(\sigma(\mathbf{\Phi}_\theta^s(Z_{t-1}))), \tag{2}
$$

where $\mathbf{\Phi}_{\boldsymbol{\theta}}^{s}(\cdot)$ is a state-specific mapping parameterized by neural networks and $\sigma(\cdot)$ is a `softmax` activation function that ensures a valid *S*-dimensional probability vector. As noted in [19], conditioning the discrete states on their preceding continuous latents (in addition to their preceding discrete states) is desirable as it allows informed transitions.

3.1.2. Switching Autoregressive Prior $p_{\theta}(Z_n|\mathcal{S}_n)$

We assume that the low-dimensional continuous latents Z_n follow a nonlinear autoregressive Gaussian prior switched by their associated discrete states $\mathcal{S}_n$. This implies a Gaussian mixture distribution for the dynamical latent space:

$$p_{\theta}(Z_t|Z_{t-\ell}, \mathcal{S}_t = s) = \text{Norm}\Big(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathbf{Z},s}(Z_{t-\ell}), \sigma_{\boldsymbol{\theta}}^{\mathbf{Z},s}(Z_{t-\ell})\Big), \qquad (3)$$

where $s \in \{1, \cdots, S\}$ and ℓ denotes a lag set (e.g., $\ell = \{1, 2\}$ for a second-order Markov model), and state-specific $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathbf{Z},s}(\cdot)$ and $\sigma_{\boldsymbol{\theta}}^{\mathbf{Z},s}(\cdot)$ are parameterized by multilayer perceptrons (MLPs) (see Table 1). In other words, we feed $Z_{t-\ell}$ to a multi-head MLP for estimation of the Gaussian parameters, e.g., $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathbf{Z},s}(Z_{t-\ell}) = \sum_{l\in\ell} \text{MLP}_{\boldsymbol{\theta}}^{s,l}(Z_{t-l})$.

Table 1. Network architectures for the nonlinear mappings in our generative model (gModel) and inference model (iModel).

gModel	$\mathbf{\Phi}_{\boldsymbol{\theta}}^{s}: \mathbb{R}^{K} \to \mathbb{R}^{S}$	$(\boldsymbol{\mu}, \sigma)_{\boldsymbol{\theta}}^{\mathbf{Z},s}: \mathbb{R}^{|\ell|\times K} \to \mathbb{R}^{K,K}$	$\mathbf{\Phi}_{\boldsymbol{\theta}}^{\mathbf{L}}: \mathbb{R}^{2K} \to \mathbb{R}^{A}$	$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathbf{X}}: \mathbb{R}^{K} \to \mathbb{R}^{3D}$
Input	$Z_{t-1} \in \mathbb{R}^{K}$	$Z_{t-\ell} \in \mathbb{R}^{|\ell|\times K}$	$h_t^{\mathcal{S}} \in \mathbb{R}^{2K}$	$Z_t \in \mathbb{R}^{K}$
1	FC $K \times K$ ReLU	FC $|\ell| \times K \times K$ ReLU	FC $2K \times K$ ReLU	FC $K \times 2K$ ReLU
2	FC $K \times K$ ReLU	AvgPool($|\ell|$)	FC $K \times A$	FC $2K \times 2K$ ReLU
3	FC $K \times S$	FC $K \times K$ ReLU		FC $2K \times 2K$ ReLU
4		FC $K \times (K + K)$		FC $2K \times 3D$
iModel	$\mathbf{\Phi}_{\boldsymbol{\phi}}^{\mathcal{S}}: \mathbb{R}^{2K} \to \mathbb{R}^{S}$		$(\boldsymbol{\mu}, \sigma)_{\boldsymbol{\phi}}^{\mathbf{Z}}: \mathbb{R}^{2K} \to \mathbb{R}^{K,K}$	
Input	$h_t^{\mathbf{X}} \in \mathbb{R}^{2K}$		$h_t^{\mathbf{X}} \in \mathbb{R}^{2K}$	
1	FC $2K \times 2K$ ReLU		FC $2K \times 2K$ ReLU	
2	FC $2K \times 2K$ ReLU		FC $2K \times 2K$ ReLU	
3	FC $2K \times S$		FC $2K \times (K + K)$	

3.1.3. Emission Model for Action Labels $p_{\theta}(L_n|\mathcal{S}_n)$

We specify a categorical emission model to determine action labels at each time point from the sequence of discrete state latents $\mathcal{S}_{1:T}$. To this end, we summarize state latents at each time point into a hidden vector $h_t^{\mathcal{S}}$ using a bidirectional LSTM recurrent network:

$$p_{\theta}(L_t|\mathcal{S}_{1:T}) = \text{Cat}\Big(\sigma\big(\mathbf{\Phi}_{\boldsymbol{\theta}}^{\mathbf{L}}\big(h_t^{\mathcal{S}}\big)\big)\Big),$$
$$h_{1:T}^{\mathcal{S}} = \texttt{B-LSTM}(\mathcal{S}_{1:T}),$$

where $\mathbf{\Phi}_{\boldsymbol{\theta}}^{\mathbf{L}}(\cdot)$ is a neural network mapping. This modeling framework implies that an action label is decided only after observing its preceding and succeeding state latents in time.

3.1.4. Emission Model for Motion Sequence $p_{\theta}(X_n|Z_n)$

We consider a Gaussian emission distribution for the observed motion instance X_t conditioned on its continuous latent Z_t. Namely, the mean of this Gaussian is a function of the continuous latent value:

$$p_{\theta}(X_t|Z_t) = \text{Norm}\big(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathbf{X}}(Z_t), \sigma^{\mathbf{X}}\text{I}\big), \qquad (4)$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathbf{X}}(\cdot)$ is a nonlinear mapping parameterized by a neural network and $\sigma^{\mathbf{X}}$ denotes observation noise.

3.2. Inference Model

As the posterior probability for this model is intractable, we use approximate variational methods in the form of amortized inference to learn the model parameters [45,46]. These methods approximate the posterior of latents $p_\theta(\mathcal{S}, Z \mid X, L)$ with a variational distribution $q_\phi(\mathcal{S}, Z \mid X)$ by maximizing the evidence lower bound (ELBO):

$$
\begin{aligned}
\mathcal{L}(\theta, \phi) &= \mathbb{E}_{q_\phi(\mathcal{S}, Z|X)} \left[\log \frac{p_\theta(X, L, \mathcal{S}, Z)}{q_\phi(\mathcal{S}, Z|X)} \right] \\
&= \log p_\theta(X, L) - \mathrm{KL}\big(q_\phi(\mathcal{S}, Z|X) \,\|\, p_\theta(\mathcal{S}, Z|X, L)\big).
\end{aligned}
\tag{5}
$$

By maximizing ELBO with respect to the parameters θ, we learn a generative model that defines a distribution over datasets pairs $p_\theta(X, L)$. By maximizing ELBO over the variational parameters ϕ, we perform Bayesian inference. Here, we assume a factorized variational distribution for the latents $\{\mathcal{S}, Z\}$:

$$
q_\phi(\mathcal{S}, Z|X) = \prod_{n=1}^{N} \prod_{t=1}^{T} q_\phi(\mathcal{S}_{n,t}|X) q_\phi(Z_{n,t}|X),
\tag{6}
$$

where its distribution parameters are estimated from input motion sequences. The graphical representation for the proposed inference model is depicted in Figure 1.

3.2.1. Variational Distributions for Discrete and Continuous Latents $q_\phi(\mathcal{S}_{n,t}|X)$, $q_\phi(Z_{n,t}|X)$

For discrete state latents $\mathcal{S}_{1:T}$, we specify a categorical variational distribution at each time point, whose parameter vector is a function of observed motion sequence. For continuous latents $Z_{1:T}$, we assume a Gaussian variational distribution at each time point, whose mean and covariance are functions of observed motion sequence. To this end, we encode each motion sequence into a hidden vector h_t^X at each time point using a bidirectional LSTM recurrent network:

$$
q_\phi(\mathcal{S}_t|X_{1:T}) = \mathrm{Cat}\Big(\sigma\big(\mathbf{\Phi}_{\boldsymbol{\phi}}^{\mathcal{S}}(h_t^X)\big)\Big),
$$

$$
q_\phi(Z_t|X_{1:T}) = \mathrm{Norm}\Big(\boldsymbol{\mu}_{\boldsymbol{\phi}}^{\mathbf{Z}}(h_t^X), \sigma_{\boldsymbol{\phi}}^{\mathbf{Z}}(h_t^X)\Big),
$$

$$
h_{1:T}^X = \texttt{B-LSTM}(X_{1:T}),
$$

where $\mathbf{\Phi}_{\boldsymbol{\phi}}^{\mathcal{S}}(\cdot)$, and $\boldsymbol{\mu}_{\boldsymbol{\phi}}^{\mathbf{Z}}(\cdot)$ and $\sigma_{\boldsymbol{\phi}}^{\mathbf{Z}}(\cdot)$ are neural network mappings that parameterize categorical and Gaussian distributions, respectively.

3.2.2. ELBO Derivation

We can derive ELBO by plugging in the generative $p_\theta(X, L, \mathcal{S}, Z)$ and variational $q_\phi(\mathcal{S}, Z|X)$ distributions from Equations (1) and (6) respectively into Equation (5) (subscript over n is dropped for brevity):

$$
\begin{aligned}
|\mathcal{L}_{n,t}(\theta, \phi)| =\ & \\
& \mathbb{E}_{q_\phi(Z_t|X)} \Big[\big\| X_t - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathbf{X}}(Z_t) \big\|_{\mathrm{F}}^2 \Big] + \\
& \mathbb{E}_{q_\phi(\mathcal{S}|X)} \Big[\texttt{CELoss}\big(\sigma(\mathbf{\Phi}_{\boldsymbol{\theta}}^{\mathbf{L}}(h_t^{\mathcal{S}})), L\big) \Big] + \\
& \sum_s q_\phi(\mathcal{S}_{t-1} = s|X) \mathbb{E}_{q_\phi(Z_{t-1}|X)} \Big[\mathrm{KL}\big(q_\phi(\mathcal{S}_t|X)||p_\theta(\mathcal{S}_t|\mathcal{S}_{t-1} = s, Z_{t-1})\big) \Big] + \\
& \sum_s q_\phi(\mathcal{S}_t = s|X) \mathbb{E}_{q_\phi(Z_{t-\ell}|X)} \Big[\mathrm{KL}\big(q_\phi(Z_t|X)||p_\theta(Z_t|Z_{t-\ell}, \mathcal{S}_t = s)\big) \Big],
\end{aligned}
$$

where $\texttt{CELoss}(\cdot, \cdot)$ denotes the cross-entropy loss function. The first two terms correspond to motion sequence reconstruction loss and action label prediction loss, respectively. The

third and fourth terms regularize discrete and continuous latent transitions, respectively. We estimate the gradients of ELBO with respect to generative and variational parameters (θ and ϕ) using a reparameterized sample from the continuous latent Z_t [47]. In the regularization terms of ELBO, the expectations over discrete latent $\mathcal{S}_t$ are easily handled by enumerating over its S possible states.

3.3. Summary of the Proposed Method

We have visualized the framework of our model in Figure 2. As shown in this figure, our model encodes an input motion sequence $\{X_1, \cdots, X_T\}$ into a sequence of hidden features $\{h_1, \cdots, h_T\}$ using a bidirectional LSTM. The resulting hidden features are fed to two separate MLPs for estimating variational distribution parameters of discrete latents $\{\mathcal{S}_1, \cdots, \mathcal{S}_T\}$ and continuous latents, $\{Z_1, \cdots, Z_T\}$. These posterior distributions are then sampled to obtain their latent values. We decode to the input motion sequence $\{\hat{X}_1, \cdots, \hat{X}_t\}$ by feeding continuous latents Z to an MLP. We also decode to the associated action labels $\{\hat{L}_1, \cdots, \hat{L}_T\}$ by feeding probability vectors of the discrete latents $\mathcal{S}$ to a bidirectional LSTM. We estimate the priors for the discrete latents $p(\mathcal{S}_t)$ and continuous latents $p(Z_t)$ from the values of sampled latents using two separate MLPs.

4. Experimental Results

We evaluated the performance of our model on two large-scale 3D skeletal motion datasets in terms of action recognition and dynamical prediction. First, we give a brief description of each skeletal dataset and the performance metrics we used throughout our experiments. Next, we introduce comparison baselines and provide implementation details of our model. Then, we describe our experimental results on each of the benchmark datasets. Finally, we conduct an ablation study to evaluate the impact of our modeling assumption. The classification and predictive performance of our model is summarized in Table 2.

Table 2. Comparison of action classification accuracy and dynamical prediction error. Our model outperformed all the baselines.

Model Dataset	Classification Accuracy (%)			Dynamical Prediction (NRMSE%)					
	Ours	P-LSTM	Ablation Ours	Ours	rSLDS	SLDS	BTMF	RKN	LSTNet
NTU (x-view)	**76.60**	70.27	69.81	**17.23**	22.45	22.19	18.81	22.64	20.22
NTU (x-sub)	**67.52**	62.93	60.74	**18.34**	23.68	25.49	21.86	25.63	24.15
Human3.6M	**78.33**	71.67	73.33	**20.76**	27.99	28.57	23.83	24.04	23.12

The best results are highlighted in bold fonts.

4.1. 3D Skeletal Datasets

We evaluated our model on a large-scale action recognition dataset, NTU RGB+D 60 [9], and a benchmark dataset for 3D human sensing in natural environments, Human3.6M [48].

NTU RGB+D 60 dataset contains 56,578 3D skeletal motion sequences for 60 action classes captured from 40 subjects and recorded by three Kinect V2 cameras concurrently from different view angles. Each motion sequence contains 3D locations of $D = 25$ skeleton joints recorded over time. Following the suggestion in [9], we split the dataset into train and test under two settings: (i) Cross-Subject (x-sub), where the subjects are split into train and test groups, yielding 40,091 and 16,487 train and test sequences, respectively, and (ii) Cross-View (x-view), where 37,646 sequences recorded from camera 2 and 3 are used for train and 18,932 sequences collected from camera 1 are used for test. We preprocessed sequences with normalization and translation following [9]. We further downsampled skeletal sequences by a factor of six (to 5 Hz) and padded all sequences to $T = 50$ time points by repetition.

Human3.6M dataset contains 3.6 million 3D human poses from 11 professional actors in 15 different scenarios (directions, discussion, eating, sitting down, greeting, taking photo, posing, making purchases, smoking, waiting, walking, sitting, phone call, walking dog, and walking together). This dataset provides accurate 3D positions of $D = 17$ body joints recorded with high-speed motion capture system at 50 Hz from two motion sequences per each actor and action label. However, only 3D pose data from 7 actors are provided (for training and validation) and the remaining poses for 4 actors are kept confidential for testing purposes in video-based pose estimation models. Therefore, in this paper for the purpose of action recognition we focus on the 7 actors from whom we have their 3D pose data and corresponding action labels. We select data from 5 actors (1, 5, 6, 7, and 8) for training and leave data from 2 actors (9 and 11) for test yielding $150 = 5 \times 15 \times 2$ motion sequences for training and $60 = 2 \times 15 \times 2$ motion sequences for test. We preprocessed sequences with normalization and translation (translating torso to origin) and further downsampled skeletal sequences by a factor of ten to 5 Hz resulting in sequences of temporal length $T = 100 - 635$.

4.2. Performance Metric

We report test set classification accuracy for quantifying the model's capacity for action recognition. We further report a histogram of dynamical state usage per action label and cluster actions based on their state-correlation matrix. In order to quantify the performance of our dynamical generative model, we compute its temporal predictive error on the test set. To this end, we predict the next time point on a test set using the generative model learned on our train set:

$$\hat{X}_{t+1} \sim p_\theta(\hat{X}_{t+1}|\hat{Z}_{t+1}),$$

where

$$\hat{Z}_{t+1} \sim p_\theta(\hat{Z}_{t+1}|Z_{t+1-\ell}, \hat{\mathcal{S}}_{t+1}),$$
$$\hat{\mathcal{S}}_{t+1} \sim p_\theta(\hat{\mathcal{S}}_{t+1}|\mathcal{S}_t, Z_t)$$
$$\mathcal{S}_t \sim q_\phi(\mathcal{S}_t|X_{1:t}),$$
$$Z_t \sim q_\phi(Z_t|X_{1:t}), \quad Z_{t+1-\ell} \sim q_\phi(Z_{t+1-\ell}|X_{1:t}),$$

where variables with hat denote predicted values and variables without hat are sampled from their posterior. In other words, latent values for the next time point are predicted from their historical values and then these predicted latent values are used for generating the next motion in the sequence. We then run inference on X_{t+1}, the actual observation at $t + 1$, to obtain Z_{t+1} and $\mathcal{S}_{t+1}$, and add them to the historical data for prediction of the next time point $\hat{X}_{t+2}$ in the same way. We repeat these steps to make predictions in a rolling manner across the test set and report their normalized root-mean-square error (NRMSE%):

$$\text{NRMSE\%} = \frac{\|X - \hat{X}\|_\text{F}}{\|X\|_\text{F}} \times 100,$$

where X and $\hat{X}$ are the ground-truth and predicted values, respectively, and $\|\cdot\|_\text{F}$ denotes Frobenius norm. Note that this metric is related to the test-set predictive log-likelihood in our case of Gaussian distributions (with a multiplicative/additive constant).

4.3. Comparison Baselines

We compared our model in terms of action classification accuracy against a state-of-the-art action recognition model, Part-Aware LSTM (P-LSTM) [9], and in terms of dynamical prediction accuracy against two established Bayesian switching dynamical models, recurrent switching linear dynamical systems (rSLDS) [20] and switching linear dynamical systems (SLDS) [18], a state-of-the-art dynamical matrix factorization method, Bayesian

temporal matrix factorization (BTMF) [49], which models higher-order linear dependencies, a state-of-the-art deep state-space model, recurrent Kalman networks (RKN) [32], which employs first-order nonlinear transitions, and a deep neural network forecasting method, long- and short-term time-series network (LSTNet) [35], which employs vector auto-regression.

The P-LSTM model divides the skeleton into five major groups of joints (torso, two hand, and two legs) and assigns a distinct LSTM cell for each body part. The outputs of these part-based cells are then concatenated to learn their common temporal patterns and are combined in the global level representation for final action classification. Similar to P-LSTM, we employ an LSTM structure for encoding motion dynamics, however, we additionally define a dynamical deep generative model to decode from these encoded features to the original motion sequences and their associated action labels. The use of LSTM structure in both P-LSTM and our method, makes P-LSTM a fair comparison baseline for the purpose of this work for evaluating the impact of our generative dynamical modeling assumption on action classification.

Dynamical baselines can be separated into two categories of "switching": rSLD and SLDS, and "non-switching": BTMF, RKN, LSTNet. rSLDS and SLDS models learn a switching dynamical model over sequential data by defining a set of discrete and continuous temporal latents similar to our method. However, they consider a first-order and linear transition model between temporal latents. On the other hand, BTMF learns dynamical transitions by defining an auto-regressive transition model on its temporal latents such that each latent is estimated from a linear combination of its preceding latents. RKN method employs a first-order transition network (parameterized by a neural network) to non-linearly map from preceding value of a temporal latent to its current value. LSTNet is a non-Bayesian deep learning-based forecasting approach which directly processes high-dimensional sequential data by employing a vector auto-regressive transition model to relate neighboring time points. Our method extends these baselines by considering a switching nonlinear auto-regressive transition model (empowered by neural networks) which is able to capture multi-modal and higher-order nonlinear dependencies in motion sequences. The aforementioned baselines share similar modeling assumptions with our method; therefore, they constitute a fair comparison.
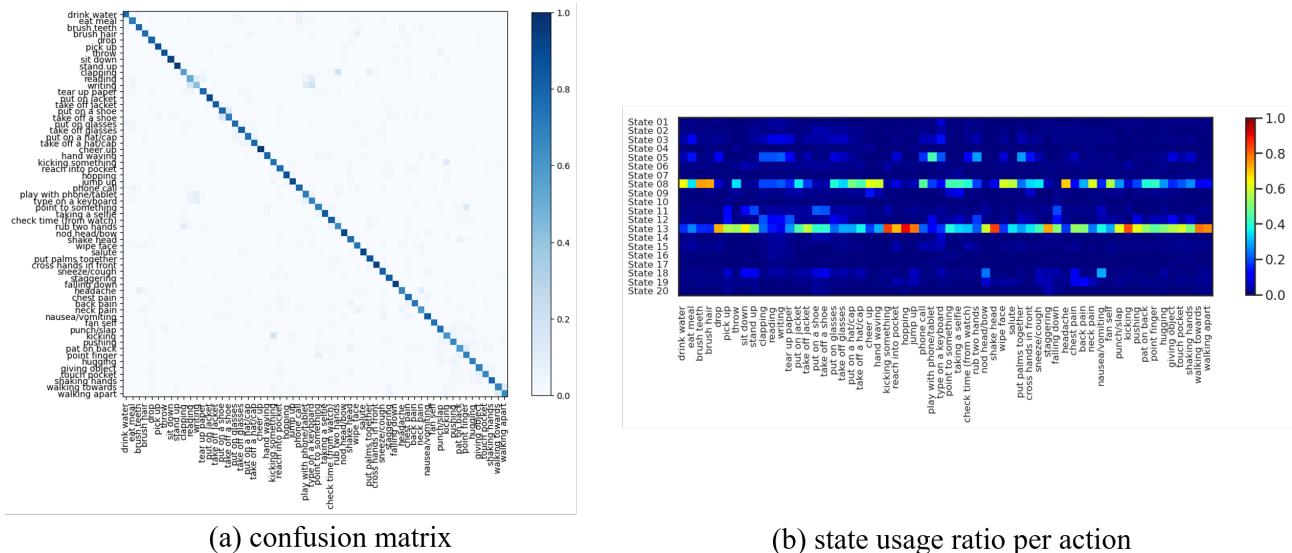
4.4. Implementation Details

We implemented our model in PyTorch v1.8 [50] and run our experiments on a TITAN Xp GPU. The network architectures for all nonlinear mappings in our model are reported in Table 1. Our model has $O(KD)$ variational and $O(S|\ell|K^2 + KD + KA)$ generative parameters. We employed Adam optimizer [51] with `lr = 0.01` and trained our models for 300 epochs. Our method took roughly 6.0 seconds per epoch with batch size of 3000 for NTU RGB+D dataset and 0.7 seconds per epoch with batch size of 1 for Human3.6M dataset. Each epoch took around 6 s.
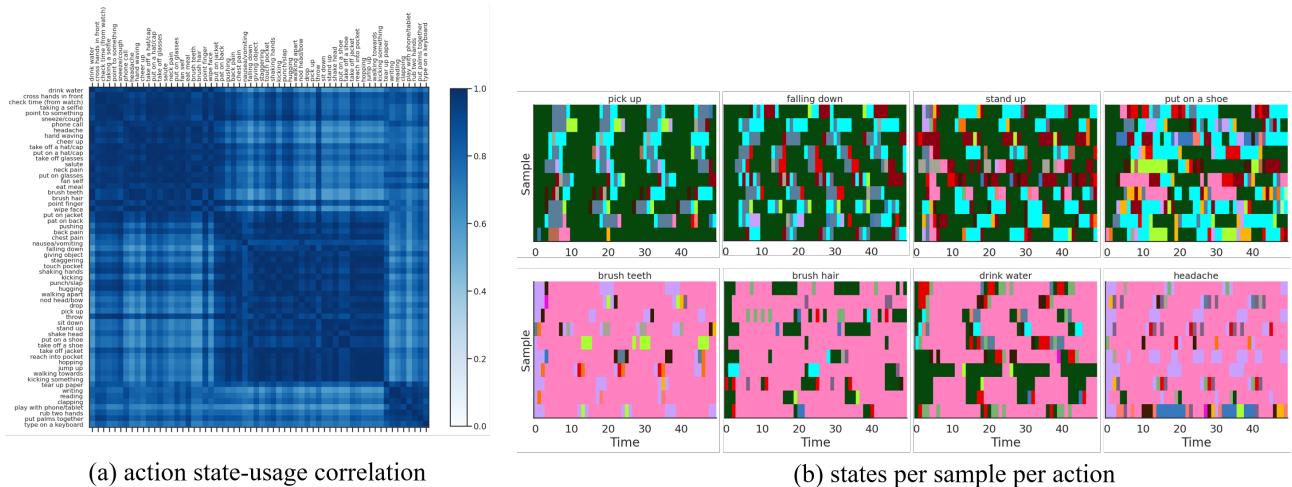
4.5. Evaluation Results on NTU RGB+D 60 Dataset

We fit our model on this dataset with $S = 20$ discrete states, $\ell = \{1, 2\}$ temporal lags (i.e., a 2nd-order Markov model), and $K = 15$ for the continuous latent dimension. We fit P-LSTM and dynamical baselines accordingly with their default settings and match their structural hyper-parameters (if applicable) for a fair comparison. As reported in Table 2, our model outperformed P-LSTM in terms of action classification accuracy in both cross-view and cross subject setup with 76.60% and 67.52% against 70.27% and 62.93%, respectively. We have visualized the confusion matrix of this classification (for x-view setup) in Figure 3a which shows difficulty in distinguishing {reading, writing, typing, playing with phone} or {clapping, rub two hands, put palms together} for instance. We have further visualized states usage shares for each action label in Figure 3b which are mainly dominated by usage of state 05, state 08 and state 13 (as major motion primitives) appearing to represent hand, upper-body and lower-body movements, respectively, for

most action labels. To further explore this, we computed action correlations in terms of their state-usage similarity, applied a spectral co-clustering algorithm [52] on the resulting correlation matrix, and visualized that in Figure 4a. This figure reveals the three major action groups based on their inferred state latents. We have visualized inferred states over time for ten randomly-selected motion sequences from each action label of {pick up, falling down, stand up, put on a shoe} and {brush teeth, brush hair, drink water, headache} in Figure 4 which confirms that similar actions share similar states as their motion primitives.

(a) confusion matrix

(b) state usage ratio per action
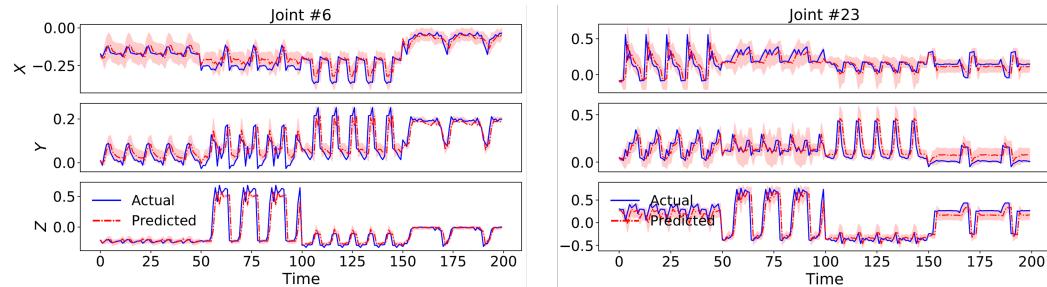
Figure 3. (**a**) Confusion matrix of action classification for x-view setup of NTU RGB+D 60 dataset. The model has difficulty in distinguishing {reading, writing, typing, playing with phone} or {clapping, rub two hands, put palms together} for instance. (**b**) State usage ratio per each action label shows predominant usage of state 05, state 08 and state 13 (as major motion primitives) appearing to represent hand, upper-body, and lower-body movements for most action labels, respectively.

(a) action state-usage correlation

(b) states per sample per action

Figure 4. (**a**) Action correlation matrix based on state-usage for x-view setup of NTU RGB+D 60 dataset, post-processed with a spectral co-clustering algorithm, reveals three major action groups. (**b**) Inferred states over time for ten randomly-selected motion sequences from each action label of {pick up, falling down, stand up, put on a shoe} and {brush teeth, brush hair, drink water, headache} confirms similar states across similar actions.
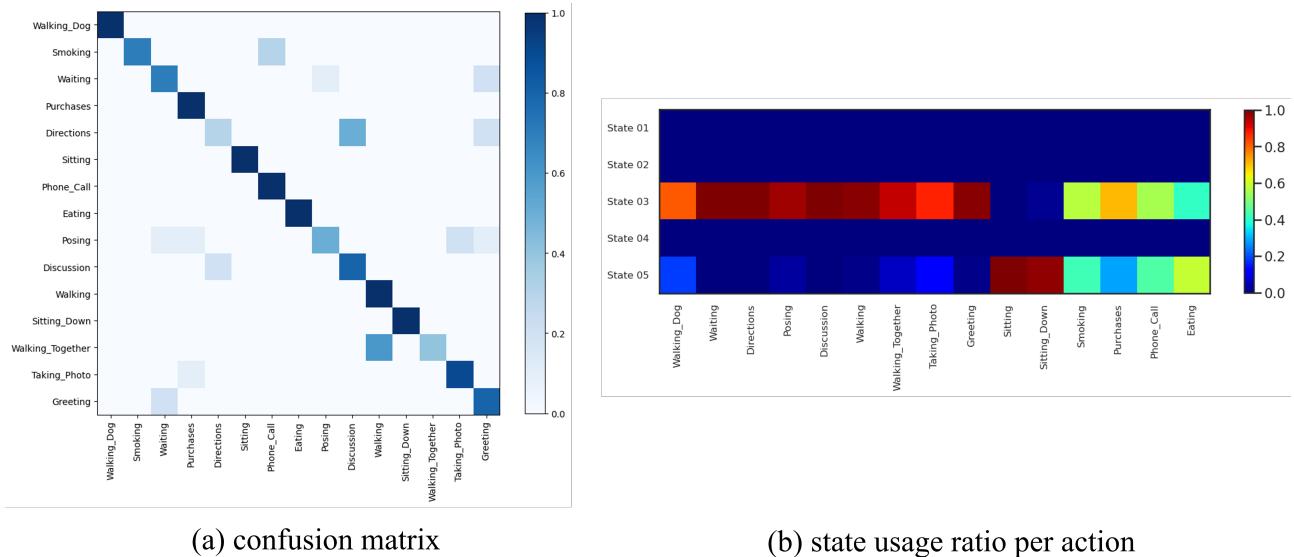
Our model also outperformed all the dynamical baselines in terms of dynamical prediction error with 17.23% and 18.34% for cross-view and cross-subject setup, respectively, by employing a 2nd-order switching nonlinear dynamical model. We have visualized test set predictions of four skeletal sequences along with their uncertainty intervals for two sample body joints in Figure 5, which indicate the capability of our model in following and predicting the actual dynamics.


Figure 5. Test set predictions of four skeletal sequences from x-view setup of NTU RGB+D 60 dataset along with their uncertainty intervals for two sample body joints.

4.6. Evaluation Results on Human3.6M Dataset
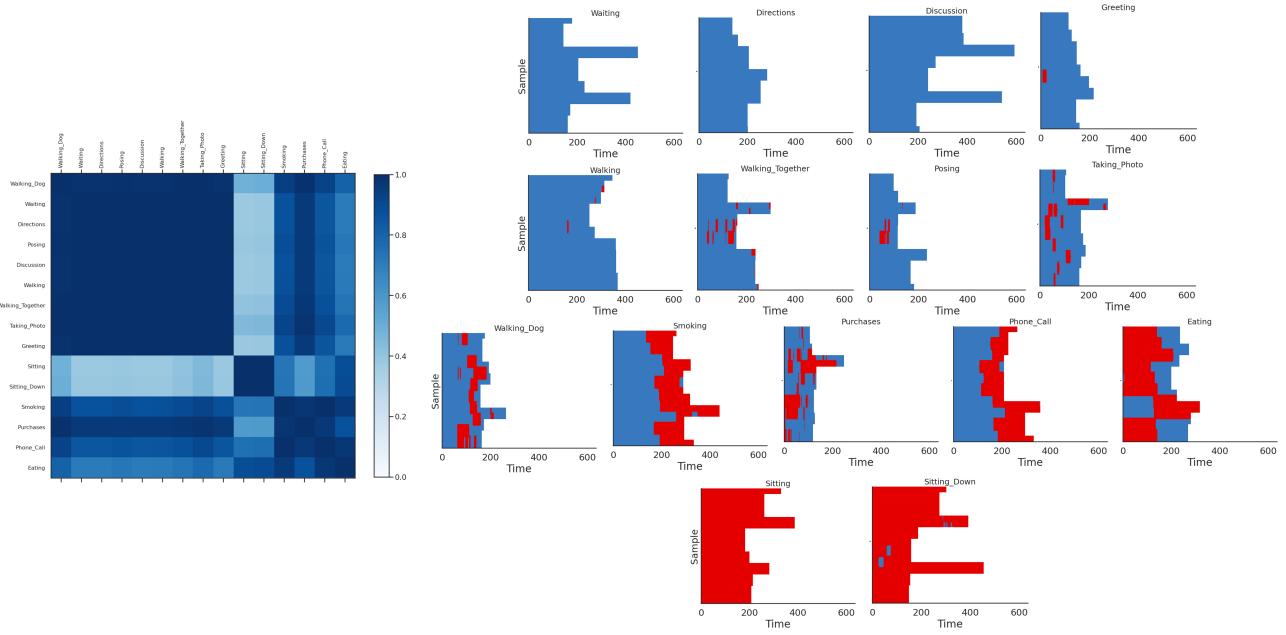
We fit our model on this dataset with $S = 5$ discrete states, $\ell = \{1, 2\}$ temporal lags, and $K = 5$ for the continuous latent dimension. We fit P-LSTM and dynamical baselines accordingly by matching their structural hyper-parameters (if applicable). As reported in Table 2, our model outperformed P-LSTM in terms of action classification accuracy with 78.33% against 71.67%, respectively. We have visualized the confusion matrix of this classification in Figure 6a which shows that the model has difficulty in distinguishing "smoking" from "phone call", "showing directions" from "discussion", "walking together" from "walking", or "waiting" and "directions" from "greeting" for instance. This is expected as these actions share very similar motion patterns and are hard to determine from pose data without any visual features. We have further visualized states usage shares for each action label in Figure 6b which are exclusive to the usage of state 03 and state 05 (as major motion primitives) appearing to represent arms and legs movements, respectively (the other three states are never used). Additionally, we computed action correlations in terms of their state-usage similarity, and applied a spectral co-clustering algorithm on the resulting correlation matrix, and visualized that in Figure 7a. This figure reveals three major action groups based on their inferred state latents. To be specific, actions of "waiting", "showing directions", "discussion", "greeting", "walking", "walking together", "posing", "taking photos" and "walking dog" are clustered together because of their dominant usage of state 03, while actions of "sitting" and "sitting down" are clustered together because of their dominant usage of state 05. On the other hand, actions of "smoking", "phone call", "eating" and "making purchases" are clustered together as they use both states (03 and 05) almost equally. While these latter actions mainly involve hands, they are mostly performed in a *sitting* posture. We have visualized inferred states over time for ten randomly-selected motion sequences from each action label in Figure 7 which confirms similar state-usage between similar actions.

(a) confusion matrix

(b) state usage ratio per action

Figure 6. (**a**) Confusion matrix of action classification for Human3.6M dataset. The model has difficulty in distinguishing "smoking" from "phone call", "showing directions" from "discussion", "walking together" from "walking", or "waiting" and "directions" from "greeting" for instance. This is expected as these actions share very similar motion patterns and are hard to determine from pose data without any visual features. (**b**) State usage histogram per each action label shows exclusive usage of state 03 and state 05 as major motion primitives which appears to represent arms and legs movements, respectively. The other three states are not utilized.
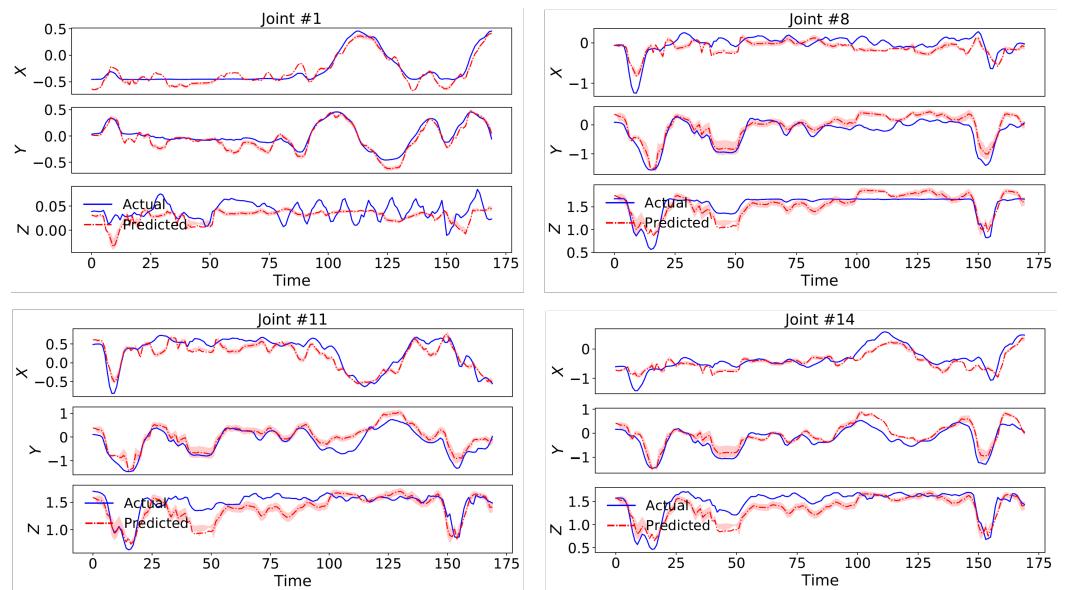
(a) action state-usage correlation

(b) states per sample per action

Figure 7. (**a**) Action correlation matrix based on state-usage for Human3.6M dataset, post-processed with a spectral co-clustering algorithm, reveals three major action groups. To be specific, actions of "waiting", "showing directions", "discussion", "greeting", "walking", "walking together", "posing", "taking photos" and "walking dog" are clustered together because of their dominant usage of state 03, while actions of "sitting" and "sitting down" are clustered together because of their dominant usage of state 05. On the other hand, actions of "smoking", "phone call", "eating" and "making purchases" are clustered together as they use both states (03 and 05) almost equally. While these latter actions mainly involve hands, they are mostly performed in a *sitting* posture. (**b**) We have visualized inferred states over time for ten randomly-selected motion sequences from each action label which confirms similar state-usage among similar actions.

Our model again outperformed all the dynamical baselines in terms of dynamical prediction error with 20.76% by employing a 2nd-order switching nonlinear dynamical model. We have visualized predictions of a test set sequence along with its uncertainty intervals for four sample body joints in Figure 8, which indicate the capability of our model in following and predicting the actual dynamics.


Figure 8. Test set predictions of a sequence in Human3.6M dataset along with its uncertainty intervals for four sample body joints, which indicate the capability of our model in following and predicting the actual dynamics.

4.7. Ablation Study

We conducted an ablation study to evaluate the impact of motion sequence reconstruction in our model on action classification accuracy. To this end, we trained a version of our model with only discrete latents which merely regresses input sequences to their action labels. The classification accuracy of this variant are reported in Table 2 which shows a significant decrease compared to the original model. We believe that the motion reconstruction term guides the model towards learning more expressive latent features (similar to encoding in auto-encoders) which are then helpful for better decoding to the action labels.

5. Conclusions

We proposed a deep switching dynamical model for action recognition and dynamical prediction in 3D skeletal motion data. Our model parsed dynamical states in the low-dimensional generative process of the data. We parameterized these low-level temporal generative models with regard to a switching deep autoregressive prior to enable multimodal and higher-order dynamical estimation. Our classification and and predictive results on two large-scale 3D skeletal datasets demonstrated the superior performance of the proposed model in comparison with the state-of-the-art methods. Specifically, our method achieved higher action classification accuracies by incorporating a dynamical generative framework in comparison with a state-of-the-art model which did not model dynamics. Our model also achieved better predictive performance (in terms of next-time point prediction on the test set) when compared to the state-of-the-art dynamical prediction models by employing a nonlinear second-order dynamical transition model. Also, the sequence of discrete latents in our method provided qualitative and quantitative interpretations about motion primitives that gave rise to each action class.

For future work, we plan to replace the LSTM structure by a graph convolutional neural network to better capture spatial correlations that are present in skeletal data by

explicitly incorporating joint-hierarchy information. There are still several limitations with our method and previous works which need to be addressed in future. A major limitation with our method and previous works in human action recognition is that (I) they are not directly generalizable to new action classes and need re-training. (II) As with other sequence modeling methods, our model can be sensitive to temporal sampling frequency. (III) Since our method processes joint locations, it is not invariant to the rotations of skeletal data and may not generalize well on globally rotated data. This is also the case with a vast body of previous works. (IV) Although our method demonstrates an improved prediction error over state-of-the-art dynamical models, its predictions may still diverge over longer horizons due to accumulation error.

References

1. Moeslund, T.; Hilton, A.; Krüger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **2006**, *104*, 90–126. [CrossRef]
2. Birch, M.C.; Quinn, R.D.; Hahm, G.; Phillips, S.M.; Drennan, B.; Fife, A.; Verma, H.; Beer, R.D. Design of a cricket microrobot. In Proceedings of the 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065), San Francisco, CA, USA, 24–28 April 2000; Volume 2, pp. 1109–1114. [CrossRef]
3. Gong, C.; Travers, M.J.; Astley, H.C.; Li, L.; Mendelson, J.R.; Goldman, D.I.; Choset, H. Kinematic gait synthesis for snake robots. *Int. J. Robot. Res.* **2016**, *35*, 100–113. [CrossRef]
4. Hoff, J.; Ramezani, A.; Chung, S.J.; Hutchinson, S. Synergistic Design of a Bio-Inspired Micro Aerial Vehicle with Articulated Wings. In Proceedings of the Robotics: Science and Systems 2016, Ann Arbor, MI, USA, 18–22 June 2016.
5. Santello, M.; Flanders, M.; Soechting, J.F. Postural hand synergies for tool use. *J. Neurosci.* **1998**, *18*, 10105–10115. [CrossRef] [PubMed]
6. Wang, P.; Li, W.; Li, C.; Hou, Y. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowl. Based Syst.* **2018**, *158*, 43–53. [CrossRef]
7. Li, Y.; Xia, R.; Liu, X.; Huang, Q. Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1066–1071.
8. Wang, H.; Wang, L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 499–508.
9. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
10. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
11. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1227–1236.
12. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with directed graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7912–7921.
13. Ackerson, G.; Fu, K. On state estimation in switching environments. *IEEE Trans. Autom. Control.* **1970**, *15*, 10–17. [CrossRef]

14. Chang, C.B.; Athans, M. State estimation for discrete systems with switching parameters. *IEEE Trans. Aerosp. Electron. Syst.* **1978**, *AES-14*, 418–425. [CrossRef]
15. Hamilton, J.D. Analysis of time series subject to changes in regime. *J. Econom.* **1990**, *45*, 39–70. [CrossRef]
16. Ghahramani, Z.; Hinton, G.E. Variational learning for switching state-space models. *Neural Comput.* **2000**, *12*, 831–864. [CrossRef]
17. Murphy, K.P. Switching Kalman Filters. Available online: https://www.cs.ubc.ca/~murphyk/Papers/skf.pdf (accessed on 15 August 2021).
18. Fox, E.; Sudderth, E.B.; Jordan, M.I.; Willsky, A.S. Nonparametric Bayesian learning of switching linear dynamical systems. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 457–464.
19. Linderman, S.; Johnson, M.; Miller, A.; Adams, R.; Blei, D.; Paninski, L. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. Available online: http://proceedings.mlr.press/v54/linderman17a/linderman17a.pdf (accessed on 15 August 2021)
20. Nassar, J.; Linderman, S.; Bugallo, M.; Park, I. Tree-Structured Recurrent Switching Linear Dynamical Systems for Multi-Scale Modeling. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
21. Becker-Ehmck, P.; Peters, J.; Van Der Smagt, P. Switching Linear Dynamics for Variational Bayes Filtering. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 553–562.
22. Farnoosh, A.; Azari, B.; Ostadabbas, S. Deep Switching Auto-Regressive Factorization: Application to Time Series Forecasting. *arXiv* **2020**, arXiv:2009.05135.
23. Sun, J.Z.; Parthasarathy, D.; Varshney, K.R. Collaborative kalman filtering for dynamic matrix factorization. *IEEE Trans. Signal Process.* **2014**, *62*, 3499–3509. [CrossRef]
24. Cai, Y.; Tong, H.; Fan, W.; Ji, P.; He, Q. Facets: Fast comprehensive mining of coevolving high-order time series. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 79–88.
25. Bahadori, M.T.; Yu, Q.R.; Liu, Y. Fast multivariate spatio-temporal analysis via low rank tensor learning. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3491–3499.
26. Yu, H.F.; Rao, N.; Dhillon, I.S. Temporal regularized matrix factorization for high-dimensional time series prediction. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 847–855.
27. Takeuchi, K.; Kashima, H.; Ueda, N. Autoregressive tensor factorization for spatio-temporal predictions. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; pp. 1105–1110.
28. Watter, M.; Springenberg, J.; Boedecker, J.; Riedmiller, M. Embed to control: A locally linear latent dynamics model for control from raw images. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2746–2754.
29. Karl, M.; Soelch, M.; Bayer, J.; van der Smagt, P. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *Stat* **2017**, *1050*, 3.
30. Krishnan, R.G.; Shalit, U.; Sontag, D. Structured inference networks for nonlinear state space models. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
31. Fraccaro, M.; Kamronn, S.; Paquet, U.; Winther, O. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3601–3610.
32. Becker, P.; Pandya, H.; Gebhardt, G.; Zhao, C.; Taylor, C.J.; Neumann, G. Recurrent Kalman Networks: Factorized Inference in High-Dimensional Deep Feature Spaces. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 544–552.
33. Farnoosh, A.; Rezaei, B.; Sennesh, E.Z.; Khan, Z.; Dy, J.; Satpute, A.; Hutchinson, J.B.; van de Meent, J.W.; Ostadabbas, S. Deep Markov Spatio-Temporal Factorization. *arXiv* **2020**, arXiv:2003.09779.
34. Chang, Y.Y.; Sun, F.Y.; Wu, Y.H.; Lin, S.D. A memory-network based solution for multivariate time-series forecasting. *arXiv* **2018**, arXiv:1809.02105.
35. Lai, G.; Chang, W.C.; Yang, Y.; Liu, H. Modeling long-and short-term temporal patterns with deep neural networks. In Proceedings of the ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 95–104.
36. Rangapuram, S.S.; Seeger, M.W.; Gasthaus, J.; Stella, L.; Wang, Y.; Januschowski, T. Deep state space models for time series forecasting. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 7785–7794.
37. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
38. Sen, R.; Yu, H.F.; Dhillon, I.S. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 4837–4846.

39. Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191. [CrossRef]

40. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 588–595.

41. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1290–1297.

42. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 143–152.

43. Muhammad, K.; Mustaqeem; Ullah, A.; Imran, A.S.; Sajjad, M.; Kiran, M.S.; Sannino, G.; de Albuquerque, V.H.C. Human action recognition using attention based LSTM network with dilated CNN features. *Future Gener. Comput. Syst.* **2021**, *125*, 820–830. [CrossRef]

44. Kwon, S. Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *Int. J. Intell. Syst.* **2021**, *36*, 5116–5135.

45. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.

46. Ranganath, R.; Wang, C.; David, B.; Xing, E. An adaptive learning rate for stochastic variational inference. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 298–306.

47. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *Stat* **2014**, *1050*, 1.

48. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339. [CrossRef]

49. Sun, L.; Chen, X. Bayesian Temporal Factorization for Multidimensional Time Series Prediction. *arXiv* **2019**, arXiv:1910.06366.

50. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. *Automatic Differentiation in Pytorch*; NeurIPS 2017 Autodiff Workshop: Long Beach, CA, USA, 2017.

51. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

52. Dhillon, I.S. Co-clustering documents and words using bipartite spectral graph partitioning. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 26–29 August 2001; pp. 269–274.