Spike Train Driven Dynamical Models for Human Actions

Michalis Raptis Kamil Wnuk Stefano Soatto
Computer Science Department,
University of California, Los Angeles, CA 90095
{mraptis,kwnuk,soatto}@cs.ucla.edu

Abstract

We investigate dynamical models of human motion that can support both synthesis and analysis tasks. Unlike coarser discriminative models that work well when action classes are nicely separated, we seek models that have fine-scale representational power and can therefore model subtle differences in the way an action is performed. To this end, we model an observed action as an (unknown) linear time-invariant dynamical model of relatively small order, driven by a sparse bounded input signal.

Our motivating intuition is that the time-invariant dynamics will capture the unchanging physical characteristics of an actor, while the inputs used to excite the system will correspond to a causal signature of the action being performed. We show that our model has sufficient representational power to closely approximate large classes of non-stationary actions with significantly reduced complexity. We also show that temporal statistics of the inferred input sequences can be compared in order to recognize actions and detect transitions between them.

1. Introduction

Analysis and synthesis of human motion is of paramount importance in human-machine interfaces, rehabilitation, security, and entertainment, just to mention a few applications. While pictorial cues convey a significant amount of information on the underlying processes, we focus on the information encoded in the temporal evolution of the data. We therefore assume that a multivariate time series has been abstracted from a person's motion, and focus on identifying models of its temporal statistics. Such a representation could be obtained from video (string of pixel intensities, orientation histograms (HOGs), joint angles of skeletal model, etc) or sensors worn by an individual. While the data extraction is by no means trivial, we focus on the second problem of learning dynamical models from the time series.

For some tasks, such as classification of distinctive motions, purely discriminative models are sufficient [9]. Some benchmark datasets can even be classified reliably taking into account as little information as local shape and optical flow in a single frame [23]. However, in situations where the temporal order of motions is significant (or perhaps the only discriminative information), or to address more subtle queries such as long-term or fine scale prediction, models with generative capability and greater representational accuracy are useful. Since the discrete multinomial state of generative models, such as Hidden Markov Models (HMMs) [31, 29, 12], experience an exponential increase in parameters as more signal history is encoded, we favor dynamic models with continuous latent variables to pursue the desired level of detail in action representation.

We propose to view human motion analysis as a blind system identification where each limb is an unknown linear dynamical system (LDS) driven by an unknown input. Intuitively, the dynamical model represents physical characteristics of an actor, such as mass and inertia, whereas the input represents the driving signal, a signature of the action. Without additional constraints this is an ill-posed problem. Traditionally, assumptions have been made that the driving input is a process with samples from some canonical distribution (typically a Gaussian). These approaches were successful at capturing observations with second-order stationary statistics, and therefore worked well for modeling quasi-repetitive actions such as walking and running [3]. However, the limitations of these models become quickly apparent when one considers more complex non-stationary sequences, e.g. Fig. 1. Our goal in this work is to be able to capture such non-stationarities in human action sequences and to reliably identify when changes between distinct actions occur.

To render the blind identification problem above well-posed we constrain the dynamics to be linear and time-invariant (our body masses do not change at the time-scale of observation), transferring all the non-stationary characteristics of the observed time series to the input. Ideally, we want a class of inputs that would serve as a signature for actions. One logical option is to assume that the input should be mostly zero, except when soliciting a

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Spike Train Driven Dynamical Models for Human Actions	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Los Angeles,Department of Computer Science,Los Angeles,CA,90095**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2010. U.S. Government or Federal Rights License

14. ABSTRACT
We investigate dynamical models of human motion that can support both synthesis and analysis tasks. Unlike coarser discriminative models that work well when action classes are nicely separated, we seek models that have finescale representational power and can therefore model subtle differences in the way an action is performed. To this end, we model an observed action as an (unknown) linear time-invariant dynamical model of relatively small order driven by a sparse bounded input signal. Our motivating intuition is that the time-invariant dynamics will capture the unchanging physical characteristics of an actor, while the inputs used to excite the system will correspond to a causal signature of the action being performed. We show that our model has sufficient representational power to closely approximate large classes of non-stationary actions with significantly reduced complexity. We also show that temporal statistics of the inferred input sequences can be compared in order to recognize actions and detect transitions between them.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

change of elementary movement. When non-zero, it should have bounded energy, to avoid embarrassing violations of elementary physics laws. This translates into the problem of performing blind identification/deconvolution under bounded energy and sparsity constraints on the input. Exploiting recent results from convex optimization and sparse representations, in Sect. 2 through Sect. 4 we develop a new algorithm for the task.

We validate our model by demonstrating the ability to accurately capture more complex actions than previous linear dynamical system approaches in Sect. 5. The compressive power of our sparse representation is also addressed. In Sect. 6, we present the application of customizable synthesis by modification of model parameters and show that our sparse representation supports segmentation and classification tasks. Through the segmentation and classification tasks, we validate our hypothesis that the input encodes signatures of individual actions.

1.1. Related Work

Understanding human actions is a critical problem that has received considerable attention in the machine learning and vision communities. Our model falls into the class of linear dynamical systems, where the task of motion modeling has been posed as a system identification problem [4, 20]. Up until now the LDS literature in human motion has assumed a stochastic input with a known distribution, which limits the representational capability to simpler regular actions. This motivates the use of switched-linear dynamical systems (SLDS), in which changes of the model parameters enhance the ability of the model to capture more complex motions [19, 15]. In [18], an SLDS approach was proposed where only the zeros of the transfer function were allowed to change across actions and an HMM was used to drive these changes. Works with a similar spirit have used switched autoregressive (SAR) systems to model videos. Video segmentation is achieved by detecting changes of the coefficients of the AR model. The identification of SAR has been addressed as a convex optimization problem by [17], and as identification of homogeneous polynomials by [27].

Yet another perspective on capturing the non-stationarity of human actions are Gaussian processes [28]. These models learn a nonlinear mapping from the observation space into a latent space and a nonlinear system in the latent space. A downside of this approach is that it does not provide information which can directly be used for classification or segmentation of the modeled motion. Physically based nonlinear temporal models have also been used to synthesize human motion [10, 11]. However, the process of concatenating "basic" controllers becomes too complex for most actions of interest.

In our case we assume a single *linear time-invariant* model. We show that by changing the assumptions on the input we increase the ability of LDS to capture complex actions and simultaneously capture useful action characteristics in the input.

Like most of the literature above we focus on designing dynamical models for actions, with no regard to how the time series is extracted. In our experiments, we use motion capture data to evaluate our approach.

2. The Underlying Dynamical System

Data observed from human actions, whether from video or motion capture, can be viewed as a multivariate time series. The core hypothesis of this work is that such multivariate time series $y(t) \in \mathbb{R}^p$ are outputs of a linear time invariant dynamical system driven by a one dimensional sparse and bounded input, $u(t) \in \mathbb{R}$. The dynamical system is defined by its system matrices $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times 1}, C \in \mathbb{R}^{p\times n}$, a state vector, $x(t) \in \mathbb{R}^n$. Above, n is the order of the LDS and p is the dimension of the observation. This model can be expressed as follows:

$$\begin{cases} \|u(\cdot)\|_{\ell_0} \leq k \\ |u(t)| \leq 1 \quad \forall\, t \\ \sum_t \|CA^tB\|_1 \leq \mu \\ x(t+1) = Ax(t) + Bu(t) \\ y(t) = Cx(t). \end{cases} \tag{1}$$

where $\|u\|_{\ell_0}$ is the number of nonzero elements in the input sequence u. We can write (1) as:

$$\begin{cases} \|U\|_{\ell_0} \leq k \\ |u_i| \leq 1 \quad \forall\, i \\ \sum_{i=0}^{N-2} \|CA^iB\|_1 \leq \mu \\ Y = \Gamma X_0 + HU \end{cases} \tag{2}$$

where i is the discrete index, N is the length of the signal, $Y = [y_0^T, y_1^T \ldots, y_{N-1}^T]^T$, $U = [u_0, u_1, \ldots, u_{N-1}]^T$, $X_0 \in \mathbb{R}^n$ is the initial condition of the system, and

$$\Gamma = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{N-1} \end{bmatrix}, H = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ CB & 0 & \ldots & 0 \\ CAB & CB & \ldots & 0 \\ \vdots & & \ddots & 0 \\ CA^{N-2}B & CA^{N-3}B & \ldots & 0 \end{bmatrix}.$$

Our representation consists of a total of 7 systems in the form of (2) that model body pose along with global position and orientation. Pose systems are learned from Euler angles grouped into 5 multidimensional time series for the body (torso, 2 arms, and 2 legs). The other 2 systems are learned from absolute 3D positions and orientation angles respectively.

The intuition behind the sparsity constraint on the input is to limit our solution space in such a way as to force as much of the stationary dynamics as possible into the system parameters. This way we hope to view our input as a triggering mechanism, a spike train sequence, that is a characteristic signature of the action. As shown in Fig. 1, the inputs found by our method given these constraints typically consist of sequences of impulses. It can thus be said that our representation interprets actions as a superposition of impulse responses.

One more aspect of our model is bounding the input. This forces variables such as the amplitude of an action into the system matrices, thus resulting in inputs that are more comparable across actions and individuals. Moreover, the unit impulse response of the system is bounded to prevent degenerate solutions due to scale ambiguity: $\|H\|_1 \to \infty$ and $\|u\|_1 \to 0$.

3. Identification with Sparse Bounded Input

Instead of seeking to minimize the one-step prediction error, as in HMMs and autoregressive models, we focus on the full simulation error:

$$\begin{aligned} \underset{U,X_0,A,B,C}{\text{minimize}} \; & \|\hat{Y} - \Gamma X_0 - HU\|_2^2 \\ \text{subject to: } & \|U\|_{\ell_0} \leq k \\ & |u_i| \leq 1 \quad , i = 0, \ldots, N-1 \\ & \sum_{i=0}^{N-2} \|CA^iB\|_1 \leq \mu \end{aligned} \tag{3}$$

where $\hat{Y}$ is the observed time series. It is well known that the minimization in (3) is NP-hard, thus we relax the problem to a weighted ℓ_1 minimization [25]:

$$\begin{aligned} \underset{U,X_0,A,B,C}{\text{minimize}} \; & \|\hat{Y} - \Gamma X_o - HU\|_2^2 + \lambda \sum_{i=0}^{N-1} w_i |u_i| \\ \text{subject to: } & |u_i| \leq 1 \quad , i = 0, \ldots, N-1 \\ & \sum_{i=0}^{N-2} \|CA^iB\|_1 \leq \mu. \end{aligned} \tag{4}$$

The form above adds a regularizer term, with λ serving as the tradeoff between accuracy of fit and sparsity.

3.1. Alternating Minimization

Our approach to solve (4) is similar in spirit to algorithms for learning dictionaries to sparsely represent images [16]. In our case, however, the dictionary is the impulse response of the linear dynamical system H.

Algorithm:

1. Select the order of the system[1]: n
2. Initialize a random sparse input U satisfying the constraint: $|u_i| \leq 1, \;\; \forall i$
3. Repeat
   (a) Given U:

   $$\begin{aligned} &\text{Identify a LDS:} \quad A,\ B,\ C,\ X_o \\ &\text{Scale } B = \min\left(1, \tfrac{\mu}{\sum_{i=0}^{N-2} \|CA^iB\|_1}\right) \cdot B \end{aligned}$$

   (b) Given X_o, Γ and H:

   $$\begin{aligned} \underset{U}{\text{minimize}} \; & \|\hat{Y} - \Gamma X_0 - HU\|_2^2 + \lambda \sum_{i=0}^{N-1} w_i |u_i| \\ \text{subject to:} \quad & |u_i| \leq 1 \quad i = 0, \ldots, N-1 \end{aligned} \tag{5}$$

For estimating the A and C matrices of the LDS we use the subspace identification algorithm for deterministic systems [26] with the constraint that A must be stable. For this purpose we adopt the method [24], which incrementally adds constraints to a quadratic program to improve the stability of the estimated system matrix. Having estimated A and C, the estimation of B and X_0 is the least-squares solution of the simulation error [7].

3.2. Enhancing Sparsity

The sparsity of the result obtained by solving a uniform weighted ℓ_1 - regularized least-squares formulation (5) can be further enhanced by incorporating an iterative reweighting scheme [6]. *Step 3(b)* of the algorithm above is thus modified as follows:

1. Initialize the weights : $w_i^{(0)} = 1, \; i = 0, \ldots, N-1$.
2. Solve the weighted ℓ_1 minimization problem

   $$\begin{aligned} U^{(l)} = \text{argmin} \; & \|\hat{Y} - \Gamma X_o - HU\|_2^2 + \lambda \sum_{i=0}^{N-1} w_i |u_i| \\ \text{subject to: } & |u_i| \leq 1 \quad i = 0, \ldots, N-1 \end{aligned}$$

3. Update the weights: $w_i^{(l+1)} = \frac{1}{|u_i^{(l)}| + \epsilon}$.
4. Terminate on convergence or when $l = l_{maxiter}$.

[1] the order of the system is verified during system identification [26].

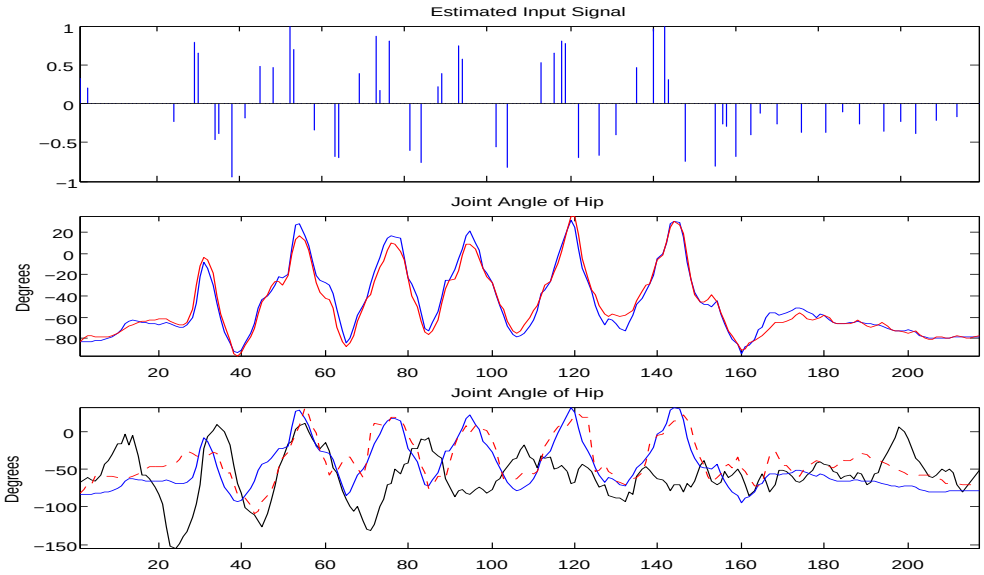


Figure 1. This figure summarizes how our model captures actions and compares the representational power of sparse input driven LDS with that of traditional stochastic LDS for non-stationary actions. The **top** plot illustrates the input to the inferred LDS that drives a person's right leg during a dance. The output of the right leg LDS corresponds to the 9 dimensions of the original joint angle time series. In the **center** plot we show that over the course of the dance we capture the joint angle of the right hip, one of the leg's dimensions, with a median error of 3.57 degrees and a mean absolute error of 4.61 degrees with a standard deviation of 3.98 degrees. The original signal is shown in blue and our corresponding synthesis is shown in red. In the **bottom** plot the synthesis result of an LDS driven by Gaussian noise is shown with a solid black line. The dashed red line shows the 5-step prediction of the same stochastic system, and the original signal appears in blue.

4. Large Scale ℓ_1 minimization

Estimating the input of a linear time-invariant (LTI) system, U, using ℓ_1 regularization is computationally intensive and becomes a challenging problem when the length of our observation is several thousand samples. However, we can reduce the computational cost significantly by exploiting the Toeplitz structure of the problem. The multiplication of a Toeplitz matrix with a vector can be performed in $O(N \log N)$ instead of $O(N^2)$. In our experiments we use the truncated Newton interior-point method proposed by Kim et al. [13], modified according to the specific constraints of our formulation. In the situation where the output is multivariate, H can be represented with p Toeplitz matrices to maintain efficiency during multiplication.

4.1. Primal and Dual Problem

In order to use the duality gap to establish convergence criteria for the minimization, we derive the dual problem here. Our initial problem is:

$$\begin{aligned} \underset{U}{\text{minimize}} \; & \|HU - \tilde{y}\|_2^2 + \lambda \sum_{i=0}^{N-1} w_i |u_i| \\ \text{subject to:} \; & |U| \preceq \mathbf{1} \end{aligned} \tag{6}$$

where $\tilde{y} = \hat{Y} - \Gamma X_0$. We change variables $\tilde{u}_i = w_i u_i$ and introduce the diagonal matrix $D = diag(\mathbf{w})$, to transform to an unweighted ℓ_1 regularized problem, where $\mathbf{w} = [w_0, \ldots, w_{N-1}]^T$. Afterward, we introduce a new variable $z \in \mathbb{R}^N$, a new equality constraint $z = HD^{-1}\tilde{u} - \tilde{y}$, and make the box constraints implicit [5].

$$\begin{aligned} \underset{-\mathbf{w} \preceq \tilde{u} \preceq \mathbf{w}, z}{\text{minimize}} \; & z^T z + \lambda \sum_{i=0}^{N-1} |\tilde{u}_i| \\ \text{subject to:} \; & z = HD^{-1}\tilde{u} - \tilde{y}. \end{aligned} \tag{7}$$

The dual function of (7) is:

$$\begin{aligned} g(\nu) &= \inf_{-\mathbf{w} \preceq \tilde{u} \preceq \mathbf{w}, z} (z^T z + \lambda \|\tilde{u}\|_1 + \nu^T (HD^{-1}\tilde{u} - \tilde{y} - z)) \\ &= \frac{\nu^T \nu}{4} - \\ &\quad \mathbf{w}^T ((D^{-1}H^T\nu + \lambda \mathbf{1})^- + (D^{-1}H^T\nu - \lambda \mathbf{1})^+) \end{aligned}$$

where $q_i^+ = \max(q_i, 0)$, $q_i^- = \max(-q_i, 0)$. Any dual feasible point ν gives a lower bound on the optimal value of the primal problem (7).

4.2. Truncated Newton Interior-Point Method

The ℓ_1 regularized least-squares problem (7) can be transformed to a convex quadratic problem, with linear in-

equality constraints.

$$\underset{\tilde{u},v}{\text{minimize}} \quad z^T z + \lambda \sum_{i=0}^{N-1} v_i \tag{8}$$

subject to:

$$z = HD^{-1}\tilde{u} - \tilde{y}; \quad -\mathbf{w} \preceq \tilde{u} \preceq \mathbf{w}; \quad -v \preceq \tilde{u} \preceq v.$$

In this part we incorporate an interior-point method for solving our convex optimization. We first define the logarithmic barrier for the bound constraints in (8):

$$\Phi(\tilde{u}, v) = -\sum_{i=0}^{N-1} log(v_i + \tilde{u}_i) - \sum_{i=0}^{N-1} log(v_i - \tilde{u}_i) - \sum_{i=0}^{N-1} log(w_i + \tilde{u}_i) - \sum_{i=0}^{N-1} log(w_i - \tilde{u}_i). \tag{9}$$

The central path consists of the unique minimizer $(x^\star(\tau), v^\star(\tau))$ of the convex function as the parameter τ varies from 0 to ∞:

$$\phi_\tau(\tilde{u}, v) = \tau \|HD^{-1}\tilde{u} - \tilde{y}\| + \tau\lambda \sum_{i=0}^{N-1} v_i + \Phi(\tilde{u}, v). \tag{10}$$

In order to minimize $\phi_\tau(\tilde{u}, v)$, the search direction is computed as an approximate solution to the Newton system, using Preconditioned Conjugate Gradient [13].

5. Experimental Evaluation

5.1. Datasets

The FutureLight action dataset [22] is a collection of 5 actions, performed with significant intra and inter-class variations: "Dance", "Jump", "Sit", "Run", and "Walk". The durations of captured actions vary from 100 to over 800 frames. We applied our learning algorithm to the full joint angle representations of all 158 samples in the dataset. In all cases we used models of order $n = 10$, with the exception of "Sit" actions which were estimated with order $n = 8$ due to the small number of available frames. We performed the deconvolution using the sparsity enhancing reweighting scheme with $\lambda = 10$ and $\epsilon = 0.005$. We use FutureLight in Sect. 5.2 and Sect. 6.3 to demonstrate accuracy and explore the supervised classification task.

To test our hypothesis about capturing action signatures in the inferred input we also obtained 6 long sequences from the CMU Motion Capture Database[2], in each of which a single actor performs several actions in succession. For instance, subject 86, sequence 3, contains smooth transitions between a number of sports related actions including walking, running, jumping, kicking, stretching, and even jump-kicking. We used the same parameter settings as in the FutureLight dataset ($n = 10$). In Sect. 6.2 we show that by taking simple statistics on the inferred input we were able to accurately classify the actions performed and localize their transitions (Fig. 2).

[2]We used ∼5000 frames from subject 86, sequences 1, 2, 3, 5, 6, 7.

5.2. Accuracy and Compression

The least requirement for a model is that it captures the statistics of the data with smaller complexity than the data itself. We show that our model achieves this task by assessing the accuracy of our reconstruction and sparsity of the inferred input.

First, we show some qualitative results of a complex non-stationary dance sequence in Fig. 1. From Fig. 1, it is clear that our model captures motions accurately where typical Gaussian noise driven LDSs of similar complexity experience a significant lack of representational power. For a more extensive evaluation, in Table 2, we report the error in representing the position (X, Y, Z) and joint angles, expressed in Euler angles, for the 5 actions in the FutureLight dataset. Further, in Table 1, we compare the mean reconstruction error for these joint angles modeled with different approaches. Our model (Y (eq. 2)) achieves the smallest reconstruction error, illustrating that it captures the signal more accurately than Gaussian noise driven LDS, even when the latter systems are given the added benefit of using information from 5 time steps in the past.

	Mean Absolute Error (^{o})
Our Model (Simulation)	**4.96**
Stoch. LDS (Simulation)	17.70
Stoch. LDS (5-step prediction)	5.29
Stoch. LDS (10-step prediction)	6.74

Table 1. Comparison with other methods.

In addition to modeling individual actions well, we observed that in the CMU sequences our model could capture successive actions and their transitions with a single dynamic system. Typically, such transitions between actions are difficult to capture and historically have even been treated as independent action classes. Results using our inference on these sequences are discussed in Sect. 6.2 and Fig. 2.

Even though synthesis is a valid method of evaluating what we capture, it is not the key goal of our model. Thus we do not focus on adding any kinematic or smoothness constraints, as is often done in graphics literature [30, 1] to generate lifelike motions.

Finally, we compute that on average, in FutureLight, 78.84% of the input signal values are zero, confirming that the inferred signal is sparse. An advantage that comes with using our sparse input LDS representation is the compressive quality of the models. For a leg, whose original N-length time series has 9 dimensions, this representation typically reduces to only 8.4% of the original size across all

	Dance XYZ	Dance Angles (^{o})	Jump XYZ	Jump Angles (^{o})	Sit XYZ	Sit Angles (^{o})	Run XYZ	Run Angles (^{o})	Walk XYZ	Walk Angles (^{o})
Mean Absolute Error	1.38	4.50	1.50	5.83	1.63	5.89	1.46	5.71	1.55	1.90
Standard Deviation of Absolute Error	1.05	3.69	1.13	5.01	1.21	4.84	1.31	4.75	1.74	2.38
Median of Absolute Error	1.16	3.66	1.28	4.64	1.37	4.72	1.15	4.61	1.08	2.38

Table 2. Representational power of our model as evaluated on the FutureLight dataset. The errors for angle measurements are in degrees. The 3D position errors are reported in the units of the motion capture data (inches scaled by a factor of 0.45).

actions (not counting the constant overhead to store the dynamical system parameters).

6. Applications

Aside from providing a concise and accurate representation of complex actions, our model allows for a number of other applications. In this section, we outline possibilities of how our model can be leveraged, and explore our hypotheses regarding the information encoded by the inferred input.

6.1. Creative Synthesis of Actions

Once a model is learned, the inputs and parameters of each dynamical system are directly available and can be controlled purposefully in ways similar to Doretto et al. for dynamic textures [8]. For example, we can change the intensity of the motion by scaling the C matrix of the system. This type of creative editing of the dynamics and input can result in interesting variations on an original action. Examples are illustrated at http://vision.ucla.edu/~mraptis/spikes

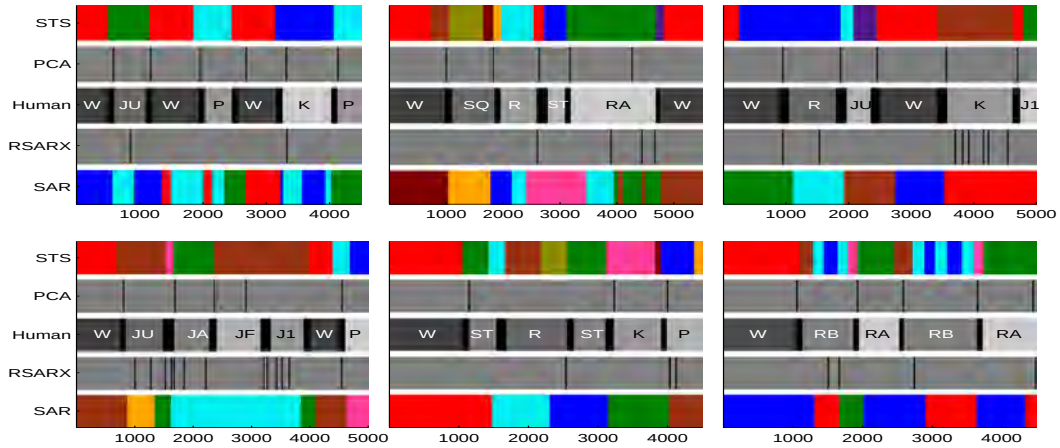
6.2. Unsupervised Action Segmentation

Having observed that our model is strong enough to capture transitions between distinct actions, we can pose the inverse problem of detecting such transitions from data. For this we use the six long sequences from the CMU dataset. Since we constrain the dynamics to be constant throughout a single sequence, finding transitions corresponds to temporal segmentation of the input. We go one step beyond simple segmentation and identify repeated patterns in the input that correspond to distinct actions. With this experiment we therefore test the hypothesis that the input signal captures signatures of observed actions.

To segment and classify the spike-trains we construct histograms of input signal intensities in a window around each frame, capturing the local statistics. Each histogram is quantized into 11 bins, equally spaced in a range from -1 to 1. To encode information from the inputs to all of the body's systems the histograms for each limb and torso are stacked to create a frame descriptor. We then use the lossy coding approach of [14] to produce an unsupervised segmentation. To encourage temporal coherence, we initially restrict merging to only neighboring segments, using a low ρ distortion parameter ($\rho = 25$). This also significantly reduces the computational cost. After convergence of the first merging the neighbor restriction is removed in order to cluster repeating actions. For this final clustering we look for the two most stable segmentations across a range of ρ values ($\rho \in [30, 130]$) and, of those, select the segmentation with the minimum ρ. Given the input identified by our algorithm, this full segmentation procedure takes approximately 1 minute to run on a 5000 frame sequence. Our results are shown in Fig. 2.

Fig. 2 also provides a comparison of our input-based segmentation results with existing algorithms for segmenting temporal data that operate directly on observed values. We compare with Barbič *et al*. [2], who proposed a change detection algorithm based on the reprojection error on the principal components computed in a sequential fashion. Their algorithm detects changes in motions relatively accurately, however it does not cluster the distinct actions. The method of Vidal [27] models the first two principal components of the data as a first order Switched Autoregressive Exogenous Model and identifies the model's coefficient recursively. Change of the model's coefficient implies change of human motion. Similarly, Ozay *et al*. [17] model the first three principal components of the data as a third order switched autoregressive model with piecewise constant coefficients. The coefficients are then clustered with K-means (with K manually selected to the optimal number for each sequence). The segmentation results of both models are shown in Fig. 2 for comparison.

As a quantitative measure of the performance of our unsupervised clustering, we compared the areas of the regions segmented with the areas of the ground truth segmentation provided by [2]. Since transitions between actions are typically smooth, (thus there are no "true" transition instants), labels assigned to regions ground truthed as transitions are not counted towards or against the classification score. Labels assigned by our algorithm were considered correct when they matched across repeated actions within a sequence and when they were unique for actions appearing only once. With this metric within-class oversegmentation does not count against us as long as it is consistent when that action class is repeated (this is observed in the "Rotate Body" actions in sequence 7). Averaging over the 6 sequences, we obtained a mean classification rate of 90.94% according to the above metric. For the SAR method the



Actions: W: walk, JU: jump up, P: punch, K: kick, SQ: squat, ST: stand, J1: jump on 1 leg, JF: jump forward, JA: jack, RB: rotate body, RA: rotate arms, R: run

Algorithms: STS: Spike Train Segmentation (our Method), PCA: [2] , HUMAN: ground truth segmentation, RSARX: [27], SAR: [17]

Figure 2. *[Best viewed in color]* Here we show the performance of our temporal segmentation/action recognition on complex CMU MoCap data (subject 86, sequences 1,2,3,5,6,7; shown left to right, top to bottom). Colors correspond to label values, thus regions marked with the same color are those that have been clustered as the same action. In sequence 2 we notice that transition regions are often identified as unique. This is due to the fact that transition regions are often smooth and do not exhibit regular statistics like their neighboring actions. In sequence 3 we notice that "walk" and "run" are confused. Likewise in sequence 5 there is confusion between "jump up", "jump forward", and "jump on one leg". In sequence 6 we oversegment "run" into 2 different labels, however neither of these are confused with any other action in the sequence. Finally the most interesting and exciting result is sequence 7. Here we have oversegmented the "rotate body" actions, but have broken them down into very regular components. We see that each instance of "rotate body" is composed of a starting transition (brown), two alternating short actions (light blue, dark blue), and an ending transition (pink).

mean classification rate was 72.27%. Our result illustrates that the distinct complex patterns of the observed data were accurately captured as patterns of the sparse input signals.

6.3. Supervised Classification

Without any supervision we deconvolve the 158 samples of the FutureLight dataset. We then classify our observations using only statistics of the input signals defining the relative pose of the actor (the actor's global position and orientation are neglected). We extract features for each sparse signal with a sliding window of length 50 and step size of 16. Within each window the features we extract include the percentage of zero elements, the percentage of successive non-zero elements that maintain the same sign, as well as the percentage of successive non-zero elements that change sign.

A dictionary is created from the extracted features with K-means clustering ($K = 12$), and each extracted window is projected onto the dictionary. At this stage, each of the 5 input signals representing an actor's body is translated into a sequence of labels. To take into account the temporal alignment of labels but also utilize support vector machines (SVM) with RBF kernel, we use the Smith-Waterman based technique described in [21] for classification. Classification performance is evaluated with a leave-one out cross validation approach, as has been used throughout the literature. We illustrate our classification results in Table 3, and compare with other methods in Table 4.

	Dance	Jump	Sit	Run	Walk
Dance	**24**	2	2		3
Jump	2	**11**		1	
Sit	1		**34**		
Run	3	3		**23**	1
Walk	5				**43**

Table 3. Confusion Matrix of Future Light Dataset. Overall mean performance 83.87%

These results confirm that basic features of the sparse input signals capture characteristics of the observed time series. In this scenario, patterns were found to be characteristic of action classes despite inference of model parameters taking place independently for each action example. We achieve reasonably good performance on this task, but do not quite match discriminative approaches that construct dictionaries directly on the multi-dimensional observation signals. However, we make the point that our model generalizes to other tasks such as segmentation and synthesis, all of which it performs in a satisfactory fashion, while the other methods lack these capabilities. We also suspect that performance on supervised classification could be further boosted by incorporating prior knowledge of the action classes into the deconvolution procedure, however we leave

	FutureLight
[21]	**98.03**
[22]	89.7
Spike Train Classification	83.63 ± 1.23

Table 4. Comparison of classification results.

this investigation for future work.

7. Conclusion

We have proposed a new and efficient alternating minimization algorithm for blind identification of linear dynamical systems driven by sparse inputs. By applying our model to a wide range of publicly available motion capture data, we have shown that this new class of models is powerful enough to capture non-stationarities of human motions. Finally, through both supervised and unsupervised segmentation and classification experiments we have demonstrated that our model is able to capture characteristic signatures of the observation in the inferred inputs. This makes it useful for analyzing sequences of various actions and applications where temporal ordering and representational accuracy are important.

Although we use motion-capture data to evaluate our dynamical models, the ultimate goal is to use these models to infer and classify time sequences of video, both at the low-level (detection and tracking) and at the high-level (action recognition).

Acknowledgments. This research is supported by ONR 67F-1080868/N00014-08-1-0414, ARO 56765-CI and AFOSR FA9550-09-1-0427. The second dataset used in this project was obtained from mocap.cs.cmu.edu. This database was created with funding from NSF EIA-0196217.

References

[1] O. Arikan, D. Forsyth, and J. O'Brien. Motion synthesis from annotations. *ACM Transactions on Graphics (TOG)*, 2003. 5

[2] J. Barbič, A. Safonova, J. Pan, C. Faloutsos, J. Hodgins, and N. Pollard. Segmenting motion capture data into distinct behaviors. In *Proc. of Graphics Interface*, 2004. 6, 7

[3] A. Bissacco, A. Chiuso, and S. Soatto. Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport. *Trans. on Pattern Analysis and Machine Intelligence*, 2007. 1

[4] A. Bissacco and S. Soatto. Classifying human dynamics without contact forces. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 2006. 2

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 4

[6] E. Candes, M. Wakin, and S. Boyd. Enhancing Sparsity by Reweighted L1 Minimization. *Journal of Fourier Analysis and Applications*, 2008. 3

[7] A. Chiuso and G. Picci. Some algorithmic aspects of subspace identification with inputs. *Int. Journal of Applied Mathematics and Computer Science*, 2001. 3

[8] G. Doretto and S. Soatto. Editable dynamic textures. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2003. 6

[9] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. of IEEE Int. Conference on Computer Vision*, 2003. 1

[10] P. Faloutsos, M. van de Panne, and D. Terzopoulos. Composable controllers for physics-based character animation. In *SIGGRAPH*, 2001. 2

[11] J. K. Hodgins and W. L. Wooten. Animating human athletics. In *SIGGRAPH*, 1995. 2

[12] N. Ikizler and D. Forsyth. Searching for complex human activities with no visual examples. *Int. J. Comput. Vision*, 2008. 1

[13] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale l1-regularized least squares. *Selected Topics in Signal Processing*, 2007. 4, 5

[14] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *Trans. on Pattern Analysis and Machine Intelligence*, 2007. 6

[15] S. O. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and inference in parametric switching linear dynamical systems. In *Proc. of IEEE Int. Conference on Computer Vision*, 2005. 2

[16] B. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. 3

[17] N. Ozay, M. Sznaier, and O. Camps. Seqential sparsification for change detection. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. 2, 6, 7

[18] V. Pavlovic, B. Frey, and T. Huang. Time-series classification using mixed-state dynamic Bayesian networks. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1999. 2

[19] V. Pavlovic, J. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *NIPS*, 2000. 2

[20] V. Pavlovic and J. M. Rehg. Impact of dynamic model learning on classification of human motion. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2000. 2

[21] M. Raptis, K. Wnuk, and S. Soatto. Flexible dictionaries for action recognition. In *Proc. of Workshop on Machine Learning for Vision-based Motion Analysis, in conjunction with ECCV*, 2008. 7, 8

[22] A. Saad, B. Arslan, and S. Mubarak. Chaotic invariants for human action recognition. In *Proc. of IEEE Int. Conference on Computer Vision*, 2007. 5, 8

[23] K. Schindler and L. Van Gool. Action snippets: How many frames does human action recognition require? In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. 1

[24] S. Siddiqi, B. Boots, and G. Gordon. A Constraint Generation Approach to Learning Stable Linear Dynamical Systems. *NIPS*, 2007. 3

[25] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996. 3

[26] P. Van Overschee and B. De Moor. Subspace Identification for Linear Systems. *Theory, implementation, applications. Kluwer Academic Publishers*, 1996. 3

[27] R. Vidal. Recursive identification of switched arx systems. *Automatica*, 44, 2008. 2, 6, 7

[28] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *Trans. on Pattern Analysis and Machine Intelligence*, 2008. 2

[29] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *Trans. on Pattern Analysis and Machine Intelligence*, 1999. 1

[30] A. Witkin and Z. Popovic. Motion warping. In *SIGGRAPH*, 1995. 5

[31] F. Xiaolin and P. Perona. Human action recognition by sequence of movelet codewords. *Proc. of 3D Data Processing Visualization and Transmission*, 2002. 1