Bio-inspired Dynamic 3D Discriminative Skeletal Features for Human Action Recognition

Rizwan Chaudhry[1], Ferda Ofli[2], Gregorij Kurillo[2], Ruzena Bajcsy[2] and René Vidal[1]
[1]Center for Imaging Science, Johns Hopkins University
[2]Tele-Immersion Lab, University of California, Berkeley

Abstract

Over the last few years, with the immense popularity of the Kinect, there has been renewed interest in developing methods for human gesture and action recognition from 3D data. A number of approaches have been proposed that extract representative features from 3D depth data, a reconstructed 3D surface mesh or more commonly from the recovered estimate of the human skeleton. Recent advances in neuroscience have discovered a neural encoding of static 3D shapes in primate infero-temporal cortex that can be represented as a hierarchy of medial axis and surface features. We hypothesize a similar neural encoding might also exist for 3D shapes in motion *and propose a hierarchy of* dynamic *medial axis structures at several spatio-temporal scales that can be modeled using a set of Linear Dynamical Systems (LDSs). We then propose novel discriminative metrics for comparing these sets of LDSs for the task of human activity recognition. Combined with simple classification frameworks, our proposed features and corresponding hierarchical dynamical models provide the highest human activity recognition rates as compared to state-of-the-art methods on several skeletal datasets.*

1. Introduction

Human activity recognition has been a topic of great research over the last several decades and has immense potential in applications such as security and surveillance, building human machine interfaces, sports training, elderly care, and entertainment.

The earliest works in modeling human motion used global representations such as skeletons and landmarks on the human body. Often cited classic work on global human motion representation is the Moving Lights Display experiment by Johansson [12] where it was shown that humans are able to recognize actions simply by the motion of the point-light displays attached to the moving subjects. 2D or 3D joint trajectories either extracted from videos or from motion capture setups have been used to extract several kinematic features such as the statistics of joint velocities or accelerations, joint angles, trajectory curvatures, etc., to represent actions [5]. Other global body feature-based approaches use hierarchical cylinders to model human limbs and torso at different scales [15] and model the 3D motion of these cylinders in space [16]. Even though approaches based on skeletal models tend to perform with high recognition rates, extracting skeletal information from 2D videos is generally very difficult, primarily because of occlusions, large variations in view-point, distortion of human shape due to clothing, and other acquisition artifacts. Motion capture system can be used to provide location of landmarks placed on the human body with high accuracy; however, such systems are usually of high complexity and require users to wear a motion capture suit with markers which can hinder the movement.

With the introduction of the Microsoft Kinect, however, a rough skeleton of a person can be easily obtained. This has resulted in renewed interest towards increased research on skeletal features for human motion representation. A number of new datasets have provided researchers with the opportunity to design novel representations and algorithms and test them on a much larger number of sequences. Recently the focus has shifted towards modeling the motion of individual joints or combinations of joints that discriminate between actions. Ofli *et al*. [19] proposed the Sequence of Most Informative Joints (SMIJ) representation, a novel and highly interpretive feature for human motion representation for skeletal data based on joint angle time series. Wang *et al*. [26] proposed a feature mining approach for computing discriminative *actionlets* from a recursively defined temporal pyramid of joint configurations.

In this paper, we propose building 3D representations of human shape and motion that are inspired by recent findings from neuroscience in the work of Yamane *et al*. [28] and Hung *et al*. [11] that attempt to find parametric models for shape space representation in the primate infero-temporal (IT) cortex. We extend their work by proposing representation of human activities as a trajectory through the 3D

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.153

shape space. In particular, we model a human activity using a hierarchy of 3D skeletal features in motion and learn the dynamics of these features using Linear Dynamical Systems (LDSs). Furthermore, instead of modeling the entire human skeleton using a single feature representation, we propose a spatio-temporal hierarchy of skeletal configurations, where each configuration represents the motion of a set of joints at a particular temporal scale. Each of these skeletal configurations is modeled as an LDS and the entire human activity is represented as a hierarchical set of LDSs. To compare different skeletal hierarchies, we develop novel discriminative metrics. From our extensive experiments we show 1) that the proposed bio-inspired features modeled using simple global LDSs already perform at par with most, if not all, state-of-the-art skeletal approaches for human action recognition, and 2) when combining these features with the proposed discriminative metric learning approach for a spatio-temporal hierarchy of LDSs, we get the highest scores for human activity recognition, to date, on several skeletal datasets.

The rest of the paper is organized as follows. In Section 2, we provide a brief description of the features proposed by Hung *et al*. [11] to describe a 3D shape space representation in primate cortex. We then briefly outline the background of LDSs and standard metrics used to compare LDSs for classification of time-series data. In Section 3, we propose a hierarchy of bio-inspired features to model human skeletal motion and model these as sets of LDSs. We then show how to compute a discriminative metric for these sets of LDSs for the purpose of classification. In Section 4 we test our proposed features and models on several datasets and conclude in Section 5.

2. Preliminaries

In this section we provide the background for our proposed bio-inspired features for 3D human action recognition. We first briefly describe the 3D static shape features used by Hung *et al*. [11] to represent the shape space in primate visual cortex. Since we are interested in moving 3D shapes, we represent a moving skeleton as a time-series of these 3D shape features and model the dynamics using a hierarchy of LDSs. We therefore provide a brief primer on time-series modeling using LDSs and present commonly used metrics for comparing LDSs.

2.1. 3D Shape Encoding in Primates

In Yamane *et al*. [28] and Hung *et al*. [11], a 3D shape is very briefly shown, by using shading and disparity cues, to a macaque monkey with their head restrained so that they cannot move. The electrical activity of several neurons is then recorded. Certain shape parameters such as surface curvature and number of medial axis components are then selectively changed according to the response of the neurons on the previously shown shape. This allows to selectively sample the 3D shape space implicitly encoded in primate cortex by observing the corresponding neuronal activity. Figure 1 from [11] shows one example of the changes in spike rate observed from one neuron as the shape space is explored by changing the parameters of the initial shape across different variations. To find the relationship between neuronal activity and 3D shape parameters, the authors in [11] showed with high confidence that the neural activity can be modeled as a non-linear function of the combination of various mathematical properties of the 3D shapes including medial-axis components such as *3D position* and *tangent directions* of points sampled along the medial axis, as well as the properties of surface fragments [28] such as *principal curvatures* and *3D orientation*. This work provides an exciting understanding of how primate brains potentially represent 3D shapes internally.

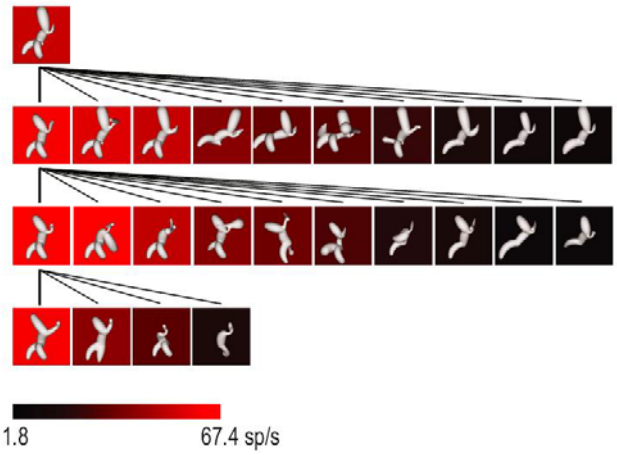

Figure 1. Various 3D shape lineages (constructed by selectively deforming the shapes across different parameters) and their corresponding neural responses in rhesus monkey cortex (see [11])

For more details about the experimental procedure and statistical validation of the authors' mathematical model of the neural 3D shape encoding space, we refer the reader to [11].

2.2. Time-Series Modeling using LDSs

Given a time series, $\{\mathbf{y}_t \in \mathbb{R}^p\}_{t=1}^T = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$, a Linear Dynamical System (LDS) models its temporal evolution using the following Gauss-Markov process:

$$\begin{aligned} \mathbf{x}_{t+1} &= A\mathbf{x}_t + B\mathbf{v}_{t+1} \\ \mathbf{y}_t &= \mu + C\mathbf{x}_t + \mathbf{w}_t. \end{aligned} \tag{1}$$

Here $\mathbf{x}_t \in \mathbb{R}^n$ represents the internal (hidden) state of the LDS at each time instant t, n represents the *order* of the LDS, $A \in \mathbb{R}^{n \times n}$ represents the *dynamics* matrix that linearly relates the states at time instants t and $t+1$, $C \in \mathbb{R}^{p \times n}$ represents the *observation* matrix that linearly transforms the internal state to the output $\mathbf{y}_t$, $\mu \in \mathbb{R}^p$ represents the

mean of the output time series. $\mathbf{v}_t \in \mathbb{R}^n$ and $\mathbf{w}_t \in \mathbb{R}^p$ correspond to the input and output noise processes usually assumed to be Gaussian with zero-mean. Specifically, $B\mathbf{v}_t \sim \mathcal{N}(0, Q)$, where $Q = BB^\top$, and $\mathbf{w}_t \sim \mathcal{N}(0, R)$, where $R = \sigma^2 I$. Given the time series $\{\mathbf{y}_t\}_{t=1}^T$, the task of computing the system parameters, $(\mathbf{x}_0, \mu, A, C, B, R)$, is referred to as system identification and several optimal [21, 23] and sub-optimal, but very efficient [9], methods have been proposed in literature.

Metrics for dynamical systems. Given a pair of LDSs, $\mathcal{M}_i = (\mathbf{x}_{0;i}, \mu_i, A_i, C_i, B_i, R_i)$ for $i = 1, 2$, existing recognition algorithms define a metric between them, $d(\mathcal{M}_1, \mathcal{M}_2)$, for the purpose of comparison. Over the years, several metrics have been proposed *e.g.* [8, 17, 25, 1, 6]. Of these, the Martin distance has been the most extensively used as it is invariant to the noise statistics as well as initial state of the dynamical system. The Martin distance compares only the parameters A and C of the dynamical models. Let $\mathcal{M}_i = (A_i, C_i)$ for $i = 1, 2$. Assuming that the systems are stable, *i.e.*, $\|A_i\|_2 < 1$, the Martin distance is defined as,

$$d_M(\mathcal{M}_1, \mathcal{M}_2)^2 = -\ln \prod_{i=1}^{n} \cos^2 \theta_i. \tag{2}$$

Here, θ_i is the i-th subspace angle between the range spaces of the infinite observability matrices O_1 and O_2 defined as

$$O_i = [C_i^\top, (C_i A_i)^\top, (C_i A_i^2)^\top, \ldots] \text{ for } i = 1, 2. \tag{3}$$

To compute the subspace angles we first solve the Sylvester equations $P_{ij} = A_i^\top P_{ij} A_j + C_i^\top C_j$ for $i, j = 1, 2$. We then compute the eigenvalues, $\{\lambda_i\}_{i=1}^{2n}$ of $\begin{bmatrix} 0 & P_{11}^{-1} P_{12} \\ P_{22}^{-1} P_{21} & 0 \end{bmatrix}$. The subspace angles, $\{\theta_i\}_{i=1}^{n}$ can then be computed as $\theta_i = \cos^{-1}(\lambda_i)$.

3. Bio-inspired Human Motion Representation

Inspired by the original work of Yamane *et al*. [28] and Hung *et al*. [11], and the promise that their shape representations present a biological encoding of shapes, we propose extracting similar features from a hierarchy of human skeletal configurations. Hung *et al*. [11] proposed a combination of medial-axis and surface features for shape representation. Since surface data is either not readily available or not of sufficient resolution and accuracy in common human activity datasets, we will focus only on the medial axis features and extend these to the time-domain for representing human activities. In the following, we explain in detail, our proposed hierarchical skeletal feature extraction procedure from each frame. We then model the dynamics of these features over the entire sequence as well as over small spatial and temporal windows using a set of LDSs. Finally, to compare different human activities for the purpose of classification, we present a method for computing discriminative metrics for these sets of LDSs.

Table 1. Human body part configurations.

No	Name	No	Name
1	HipJoint [HJ]		
2	RightUpLeg [RUL]	3	LeftUpLeg [LUL]
4	RightLeg [RL]	5	LeftLeg [LL]
6	RightFoot [RF]	7	LeftFoot [LF]
8	Spine1 [S1]	9	Spine2 [S2]
10	Neck [N]		
11	RightArm [RA]	12	LeftArm [LA]
13	RightForeArm [RFA]	14	LeftForeArm [LFA]
15	RightFullLeg (RUL + RL + RF) [RFuL]		
16	LeftFullLeg (LUL + LL + LF) [LFuL]		
17	LowerBody (RFL + LFL + HJ) [LB]		
18	RightFullArm (RA + RFA) [RFuA]		
19	LeftFullArm (LA + LFA) [LFuA]		
20	UpperBodyAndArms (S2 + RFuA + LFuA) [UBA]		
21	BackAndNeck (HJ + S1 + S2 + N) [N]		
22	FullUpperBody (BN + UBA) [FUB]		
23	FullBody (FUB + LB) [FB]		

3.1. Hierarchical Medial-Axis Template Models for Human Skeletal Configurations

Figure 2(a) provides an example of the human skeleton structure from the Berkeley MHAD [20]. Following [11], we divide the human skeleton into several topological parts such as chains, single X/Y junctions, and double X/Y junctions. In [11], the authors divided a 3D shape into all possible topological parts. However the shapes under investigation in [11] were much simpler and had at most two to eight axial components. The human skeleton on the other hand has many more axial components and enumerating all possible topological parts becomes computationally prohibitive. We therefore propose using a fixed hierarchy of semantically meaningful body parts and extract the medial-axis features proposed in [11] from each of these parts. Figure 2(b) illustrates the skeletal part hierarchy which is explicitly defined in Table 1. As we can see, each of these body parts can be categorized as a chain, an X/Y junction or a double X/Y junction.

Once the skeletal part hierarchy is defined, we describe the features extracted from each of these parts. Note that before extracting any features, all the 3D joint coordinates are transformed from the world coordinate system to a person-centric coordinate system by placing the hip joint at the origin. Another alignment step could potentially be performed to rotate the world coordinate frame to a person-centric frame; however, in general the sequences in most datasets are taken from the same viewpoint. This is especially true for the Kinect where a single camera is used to acquire data for skeletal feature extraction.

We first describe the process of extracting features from

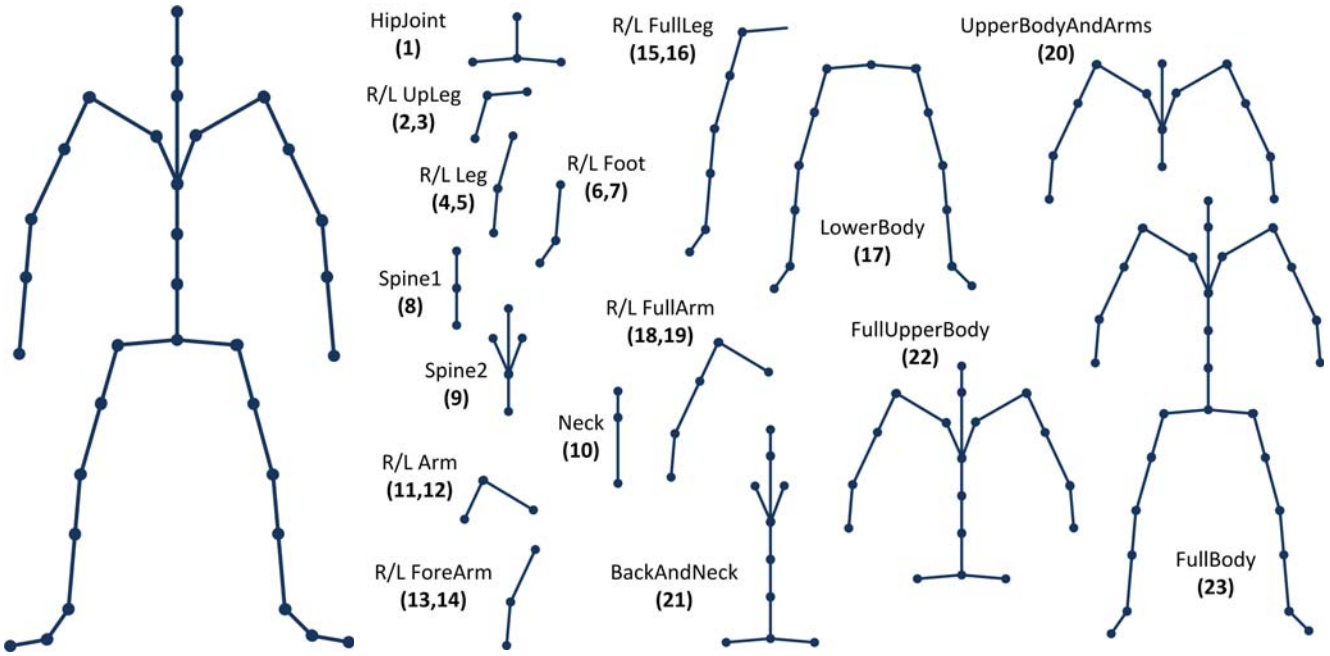

Figure 2. (a) Joint locations and connectivity for the Berkeley MHAD. (b) Different skeletal part configurations as presented in Table 1.

a chain which is also the building block for X/Y and double X/Y junctions. Given the transformed 3D coordinates of a chain, following [11], we sample 21 equi-distant points along the chain and compute the tangent direction at each of these points. We also compute a simple 3D extension of the shape context feature [4] using these points. There have been several works in extending the shape context to 3D, *e.g.*, [13, 10, 27]. However we use a very simple histogram binning procedure, whereby we divide the sphere into 4 equally spaced longitudinal sections and 8 equally-spaced latitudinal sections for a total of 32 quantized directions. At each of the 21 equally spaced points, we compute the normalized histogram of directions to all other points in the chain. This results in a $21 \times 32 = 672$ dimensional 3D shape-context feature for the entire chain. Therefore, from each chain, we extract a $21 \times 3 = 63$-dimensional feature for 3D position, a $21 \times 3 = 63$-dimensional feature for 3D tangent direction and a 672-dimensional shape context feature. The full 798-dimensional feature represents the spatial configuration of the skeletal part at each time-instant. Figure 3 illustrates these points for a chain with two links and the extracted features. For X- and Y-junctions and double X/Y junctions, we compute 21 equidistant points along each chain constituting the junctions. Similarly, we compute the 3D positions and tangent directions at each point. The shape context features at each sampled point, however, are computed by using all the points of the entire junction.

A note about shape context features. In general, shape context is used for shape matching and computing the best alignment of two similar shapes when point correspondences are unknown. In the case of skeletal configurations, we know exactly the correspondences between the points on any two skeletons and hence, to compare two shape context features, we do not need to find the configuration that

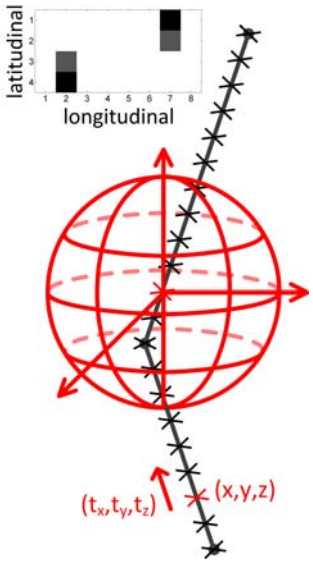

Figure 3. Skeletal feature extraction procedure for a chain illustrating equi-distant point sampling, tangents and 3D direction quantization for shape context computation.

minimizes an alignment distance. Instead we treat shape context features as simple Euclidean vectors and compute the difference between two shape context features by simple Euclidean subtraction.

A skeletal action sequence, therefore, is represented as several time series of skeletal medial axis features (position, tangents and shape context features). Each time series corresponds to features extracted from the part descriptions in Table 1. We further sub-divide the time series into several temporal scales starting from the entire time-series data in a particular sequence to smaller and smaller equal-sized temporal parts. This results in a fixed number of individual time series corresponding to different body part configurations and different temporal extents. We denote by $Y_{(k,t,h)}$, the feature time series extracted from body part configuration $k \in \{1, \ldots, K\}$ and the h-th temporal window at normalized temporal scale τ^{-1}, where $\tau = 2^t, t \in \{0, 1, 2, \ldots T\}$ and $h \in \{1, \ldots, 2^t\}$. We then model each individual feature time series using an LDS and learn the corresponding system parameters $\mathbf{M}_{(k,t,h)}$ as outlined in Section 2.2. Hence, the set of feature time series extracted from a skeletal sequence is represented as the set of LDSs, $\{\mathbf{M}_{(k,t,h)}\}_{k=1,\ldots,K}^{t=0,\ldots,T, \quad h=1,\ldots,2^t}$. Out of these $K \times \left(2^{T+1} - 1\right)$ sets of LDS parameters, $M_{(k=K,t=0,h=1)}$ corresponds to the parameters of the full global feature time series extracted from the full body (which corresponds to a double X/Y junction) for the entire length of the skeletal sequence.

3.2. Discriminative Metric Learning for Sets of LDSs

In the previous section, we proposed modeling human actions recorded by motion capture data by using a set of medial-axis feature time series. We can compare two global medial-axis feature time series by computing the metric between their corresponding system parameters as outlined in Section 2.2. In a similar fashion, we can compute the metrics between all corresponding (k, t, h) pairs of system parameters and define a new similarity metric between two sets of LDSs, $\{\mathbf{M}^1_{(k,t,h)}\}_{k=1,\ldots,K}^{t=0,\ldots,T, \quad h=1,\ldots,2^t}$ and $\{\mathbf{M}^2_{(k,t,h)}\}_{k=1,\ldots,K}^{t=0,\ldots,T, \quad h=1,\ldots,2^t}$ by combining the values of the metrics computed between each $\mathbf{M}^1_{(k,t,h)}$ and $\mathbf{M}^2_{(k,t,h)}$. We propose using Multiple Kernel Learning (MKL) [3, 24] to learn a set of optimal weights, $\alpha_{(k,t,h)}$, in a supervised fashion such that the weighted linear combination of kernels computed individually from each part configuration and temporal extent gives the best action recognition performance. For conciseness, we define $\mathbf{M}^i \doteq \{\mathbf{M}^i_{(k,t,h)}\}_{k=1,\ldots,K}^{t=0,\ldots,T, \quad h=1,\ldots 2^t}$.

For a two-class problem, the kernel between the set of system parameters from two skeletal sequences can then be written as,

$$k(\mathbf{M}^i, \mathbf{M}^j) = \sum_{k=1}^{K} \sum_{t=0}^{T} \sum_{h=1}^{2^t} \alpha_{(k,t,h)} k_{\text{LDS}}(\mathbf{M}^i_{(k,t,h)}, \mathbf{M}^j_{(k,t,h)}). \tag{4}$$

Given the class labels, L_i for each $\mathbf{M}^i$ in the training set, MKL learns the optimal values of α for the best classification performance on the training set. Using these values of α, a full Kernel-SVM classifier is trained on the training set and used to classify a new sequence with the weighted kernel. Several variations of the MKL algorithm have been proposed with various regularization choices for the weight vector α (See [2] for a review). Furthermore, there are several extensions to the multi-class case. We will use SimpleMKL [22] that simultaneously optimizes over the sum of all one-vs-all or one-vs-one classifiers and uses the same set of α weights in all classifiers.

MKL has generally been shown in practice to provide excellent classification rates when using different types of features to model the same phenomenon. Furthermore, the weights can be used to reason about the relative importance of some features for the purpose of classification. As we show in Section 4, we get superior human activity recognition performance when combining multiple body part configurations across several temporal scales as opposed to only using the medial-axis features extracted from the full body and learning the dynamics over the entire time series.

4. Experiments

We now show experimental results for human activity recognition in skeletal data using our proposed dynamic medial-axis features. We evaluate the performance of our approach when using only one global LDS, $M_{(k=K,t=0,h=1)}$, for the entire medial-axis feature time series for the full human body as well as when combining the features across several body configurations and several temporal scales.

4.1. Datasets

We report action recognition results on the Berkeley MHAD [20], HDM05 [18] and MSR Action3D [14] datasets. A brief description of these datasets follows:

Berkeley Multimodal Human Action Database (Berkeley MHAD). This dataset contains 11 actions performed by 12 subjects with 5 repetitions of each action, yielding a total of 659 action sequences (after excluding an erroneous sequence). The motion capture data was recorded at 480 fps and the action lengths vary from 773 to 14565 frames (corresponding to approximately 1.6 to 30.3 seconds). In our experiments, we used 7 subjects (384 action sequences) for training and 5 subjects (275 action sequences) for testing. The set of actions consisted of *jump, jumping jacks,*

bend, punch, wave one hand, wave two hands, clap, throw, sit down, stand up, and *sit down/stand up*.

Motion Capture Database HDM05. From the popular HDM05 database [18] we arbitrarily selected 11 actions performed by 5 subjects. In this dataset, subjects performed each action with various number of repetitions, resulting in 251 action sequences in total. The motion capture data, which was captured with the frequency of 120 Hz, also includes the corresponding skeleton data. The duration of the action sequences ranges from 121 to 901 frames (corresponding to approximately 1 to 7.5 seconds). In our experiments, we used 3 subjects (142 action sequences) for training and 2 subjects (109 action sequences) for testing. The set of actions consisted of *deposit floor, elbow to knee, grab high, hop both legs, jog, kick forward, lie down floor, rotate both arms backward, sneak, squat,* and *throw basketball*.

MSR Action3D Database. Finally, we also evaluated the action recognition performance on the MSR Action3D dataset [14] consisting of the skeleton data obtained from a depth sensor similar to the Microsoft Kinect with 15 Hz. Due to missing or corrupted skeleton data in some of the action sequences available, we selected a subset of 17 actions performed by 8 subjects, with 3 repetitions of each action. The subset consisted of 379 action sequences in total, with the duration of the sequences ranging from 14 to 76 frames (corresponding to approximately 1 to 5 seconds). We used 5 subjects (226 action sequences) for training and 3 subjects (153 action sequences) for testing. The set of actions included *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, forward kick, side kick, jogging, tennis swing,* and *tennis serve*.

4.2. Global LDS Models

We first perform several baseline experiments using features extracted using the entire skeleton across the full temporal extent. This allows us to compare how well each of the features - position, tangents and shape context as well as their combination - performs on the entire skeleton. For each of the datasets above, we compute the global medial-axis feature time series and identify the parameters of an order $n = 5$ system to model the linear dynamics. We separately compute the LDS parameters of only the 3D coordinates, the 3D tangent directions and the 3D shape context features as well as their concatenation to determine which feature is the most discriminative. We then use the hybrid Martin metric for comparing LDSs. The hybrid Martin metric is a weighted combination (we use equal weights) of the Martin distance as defined in Section 2 and the Euclidean difference of the temporal means and is generally shown to perform much better than the Martin distance alone (See [7] for more details).

Table 2. Activity recognition rates for global LDS models for different medial-axis features.

	Berkeley MHAD		HDM05		MSR Action3D	
Method	1-NN	SVM	1-NN	SVM	1-NN	SVM
3D position	96.73	99.27	93.58	91.74	73.33	80.00
Tangents	95.64	98.91	91.74	88.07	76.67	84.44
Shape	87.27	93.09	88.99	82.57	75.56	83.33
All	97.09	99.27	82.66	90.83	78.33	83.89

Table 3. Classification results for several baseline representations described in [19]. SMIJ - Sequence of the Most Informative Joints, HMIJ - Histograms of the Most Informative Joints, HMW - Histogram of Motion Words, LDS - Linear Dynamical System modeling of joint angle trajectories.

	Berkeley MHAD		HDM05		MSR Action3D	
	1-NN	SVM	1-NN	SVM	1-NN	SVM
SMIJ	78.91	94.18	80.73	84.40	24.18	29.41
HMIJ	72.73	82.91	80.73	82.57	26.14	29.41
HMW	70.91	81.09	78.90	78.90	21.57	32.68
LDS	69.45	82.18	72.48	76.15	43.14	47.06

Table 2 shows the recognition rate when using 1-NN and SVM for classification on all three datasets. We can observe that although 3D position alone also performs the best in some cases, overall, using all three types of features performs better than using each feature separately. As a result, we get almost perfect activity recognition on the Berkeley MHAD by using our full body medial-axis features when modeled using LDS.

To compare these results with the state-of-the-art algorithms on these datasets, in Table 3, we have reproduced the results in [19] for modeling joint angle variations of the human skeleton as the person performs different activities. The results in the last row of Table 3 correspond to modeling the global dynamics of joint angle trajectories for each dataset. Joint angle trajectories have traditionally been modeled using dynamical systems and HMMs to represent human activities. We can see that joint angle trajectories alone do not capture adequate information to be discriminative enough about the activity being performed, whereas due to the sampling of points along the human skeleton, our medial-axis feature representation captures much more information and provides much higher recognition rates on all three datasets.

For further comparison to state-of-the-art methods on the MSR Action3D dataset, we have also reproduced the results in [26] in Table 4. All the methods except for the one proposed in [26] performed worse than our global LDS model on the skeletal features. These comparisons provide strong support in favor of our proposed features for human skeletal action recognition.

4.3. Discriminative Hierarchy of LDSs

As proposed in Section 3.2, we now use MKL to learn the optimally discriminative weights for the entire set of body part configurations across several temporal scales instead of only using the entire human body and all the frames

Table 4. Performance of state-of-the-art methods on the MSR Action 3D dataset as shown in [26].

Method	Accuracy
Recurrent Neural Network	42.5
Dynamic Temporal Warping	54
Hidden Markov Model	63
Action Graph on Bag of 3D Points	74.7
Actionlet Ensemble [26]	88.2

per sequence. For the Berkeley MHAD dataset, we divide the skeleton into a hierarchy of 23 body parts as proposed in Table 1. We use 5 different temporal scales by dividing each video equally into 2^t temporal parts, where $t \in \{0, 1, 2, 3, 4\}$. This gives a total of 31 temporal windows of different sizes spanning a range of frames from the entire video to $\frac{1}{16}$ of the video. According to the formulation in Section 3.2, this results in a total of $31 \times 23 = 713$ different body part feature time series. We learn the system parameters of an order $n = 5$ LDS for each of these time series and represent an action sequence by the set of LDS parameters. Given training labels, we then use MKL to learn the optimal weights for an RBF kernel constructed using the hybrid Martin distance between the LDS parameters of corresponding body-parts at the same temporal scale. Once these weights are learnt, we use them in a regular kernel SVM for classification on the test set.

Figure 4 shows the weights computed using SimpleMKL [22] for different body-part configurations and temporal locations for the Berkeley MHAD dataset. As we can see, the highest number of positive weights are associated to the features extracted from the full body (index 23) which is consistent with the good performance of the full body features in the previous section. However, the largest weight is associated to the lower body (index 17) feature at a temporal scale of $\frac{1}{2}$ and the LeftForeArm (index 14) at a temporal scale of $\frac{1}{4}$. Having observed that, it remains difficult to associate a non-qualitative reason as to why a certain body-part configuration at a particular temporal scale and temporal location is discriminative. However when these learned feature weights are used to perform classification using kernel SVM on the test data, we get 100% correct classification on the Berkeley MHAD as can be seen from Table 5.

We similarly tested our proposed approach on the HDM05 and MSR Action3D datasets. The recognition results are provided in Table 5. Since the frame rate of the MSR Action 3D dataset is only 15 frames/sec and some videos have fewer than 15 frames, it is not possible to extract sufficient time-series data for LDS parameter estimation for smaller temporal extent sub-sequences. We therefore up-sample the skeleton data to 120 Hz (equal to that of HDM05) before extracting medial axis features from this dataset. As we can see, compared to Tables 3,4, and the global LDS approach in Table 2, we achieve the best possible results using our proposed discriminative LDS parts

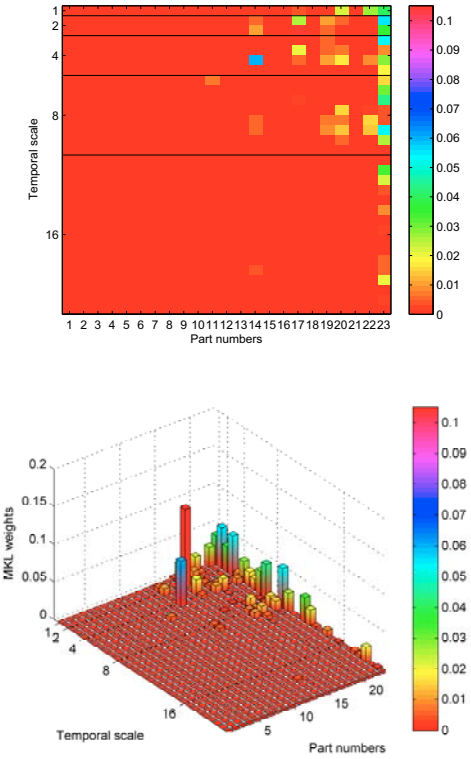

Figure 4. MKL weights corresponding to body part configurations and temporal extents for the Berkeley MHAD dataset. Top: scaled image of weight matrix. Bottom: 3D bar plot of weight matrix. The x-axis corresponds to the part numbers from Table 1 and the y-axis corresponds to the temporal scale and the particular temporal window number. For example, when dividing the sequence into 2 parts, there are 2 temporal windows and the figures show the weight corresponding to each of these windows. The corresponding recognition rate is 100%.

Table 5. Action recognition performance for all three datasets when learning our proposed discriminative LDSs for all body-part configurations and temporal scales.

Dataset	Accuracy
Berkeley MHAD	100
HDM05	98.17
MSR Action3D	90.00

approach. We also get better results on the MSR Action3D dataset than those in the state-of-the-art Actionlet Ensemble method in [26].

5. Conclusions

In this paper, we have leveraged the recent advances in the area of static shape encoding in the neural pathway of primate cortex, and proposed new bio-inspired features for human activity recognition in skeletal data. We have extended the neural static shape encoding features to represent moving shapes such as humans by using a discriminative set of LDSs. Our experiments on several human activity skeletal datasets have shown very successful results. This

provides strong evidence in favor of the efficacy of these features in general, and combined with our proposed hierarchical extension to the spatio-temporal domain along with the dynamical modeling as sets of LDSs, as excellent representations for shapes in motion. Apart from being very useful for human activity recognition in 3D data sources such as the Kinect, our study might also provide some impetus to the neuroscience community where our proposed dynamical models could be used to determine neurological models for moving shapes.

Acknowledgments. We would like to thank C.-C. Hung and C.E. Connor for helping us in fully understanding the neural shape encoding in [11]. This work was supported in part by the European Research Council grant VideoWorld as well as the grants NSF 0941362, NSF 0941463, NSF 0941382 and ONR N000141310116.

References

[1] B. Afsari, R. Chaudhry, A. Ravichandran, and R. Vidal. Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic visual scenes. 2012.

[2] M. G. E. Alpaydin. Multiple kernel learning algorithms. 12:2211–2268, 2011.

[3] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. 2004.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. 24 (24):509–521, 2002.

[5] C. Cedras and M. Shah. Motion-based recognition: A survey. *Image and Vision Computing*, 13 (2):129–155, 1995.

[6] A. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. volume 1, pages 846–851, 2005.

[7] R. Chaudhry and R. Vidal. Recognition of visual dynamical processes: Theory, kernels and experimental evaluation. Technical Report 09-01, Department of Computer Science, Johns Hopkins University, 2009.

[8] K. D. Cock and B. D. Moor. Subspace angles and distances between ARMA models. *System and Control Letters*, 46(4):265–270, 2002.

[9] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. 51(2):91–109, 2003.

[10] K. S. Huang and M. M. Trivedi. 3D shape context based gesture analysis integrated with tracking using omni video array. 2005.

[11] C.-C. Hung, E. T. Carlson, and C. E. Connor. Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, 74 (6):1099–1113, 2012.

[12] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 1414 (2):201–211, 1973.

[13] M. Koertgen, G. Parl, M. Novotni, and R. Klein. 3D shape matching with 3D shape contexts. In *Seventh Central European Seminar on Computer Graphics*, 2003.

[14] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 9 –14, june 2010.

[15] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 200 (1140):269–294, 1978.

[16] D. Marr and L. Vaina. Representation and recognition of the movements of shapes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 214 (1197):501–524, 1982.

[17] A. Martin. A metric for ARMA processes. *IEEE Trans. on Signal Processing*, 48(4):1164–1170, 2000.

[18] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation of MOCAP database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[19] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. In *International Workshop on Human Activity Recognition from 3D Data*, 2012.

[20] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. MHAD: A comprehensive multimodal human action database. In *IEEE Workshop on Applications of Computer Vision*, 2013.

[21] P. V. Overschee and B. D. Moor. N4SID : Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica, Special Issue in Statistical Signal Processing and Control*, pages 75–93, 1994.

[22] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[23] R. Shumway and D. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.

[24] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. 7 (1):1531–1565, 2006.

[25] S. Vishwanathan, A. Smola, and R. Vidal. Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. 73(1):95–119, 2007.

[26] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. 2012.

[27] D. Xiao, D. Zahra, P. Bourgeat, P. Berghofer, O. A. Tamayo, C. Wimberley, M. C. Gregoire, and O. Salvado. An improved 3d shape context based non-rigid registration method and its application to small animal skeletons registration. *Computerized Medical Imaging and Graphics*, 34(4):321 – 332, 2010.

[28] Y. Yamane, E. T. Carlson, K. C. Bowman, Z. Wang, and C. E. Connor. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, 11 (11):1352–1360, 2008.