Correction: Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group

In this paper, we have reported the classification accuracy for MSR-Action3D dataset under two experimental protocols: the subset-based protocol of [7] and the full dataset-based protocol of [19]. Recently, a bug has been found in the code that was used to compute the accuracy under the protocol of [7]. After fixing this bug, the average classification accuracy values reported in Table 2 of this paper have increased. The below tables show the earlier reported results and the new results. Note that the proposed skeletal representation still performs better than joint positions, joint angles and body part locations. However, its performance is slightly lower (0.46%) when compared to relative joint positions.

We thank Haomiao Ni from South China University of Technology for pointing out the bug in the code.

This version of the paper is a corrected version with all the corrections highlighted in red.

Recognition rates reported in the paper for various skeletal representations on MSR-Action3D dataset using the protocol of [7]

Dataset	JP	RJP	JA	BPL	Proposed
AS_1	91.65	92.15	85.80	83.87	95.29
AS_2	75.36	79.24	65.47	75.23	83.87
AS_3	94.64	93.31	94.22	91.54	98.22
Average	87.22	88.23	81.83	83.54	**92.46**

New recognition rates for various skeletal representations on MSR-Action3D dataset using the protocol of [7])

Dataset	JP	RJP	JA	BPL	Proposed
AS_1	93.36	95.77	84.51	90.30	94.72
AS_2	85.53	86.90	68.05	83.91	86.83
AS_3	99.55	99.28	96.17	95.39	99.02
Average	92.81	**93.98**	82.91	89.87	93.52

Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group

Raviteja Vemulapalli, Felipe Arrate and Rama Chellappa
Center for Automation Research
UMIACS, University of Maryland, College Park.

Abstract

Recently introduced cost-effective depth sensors coupled with the real-time skeleton estimation algorithm of Shotton et al. [16] have generated a renewed interest in skeleton-based human action recognition. Most of the existing skeleton-based approaches use either the joint locations or the joint angles to represent a human skeleton. In this paper, we propose a new skeletal representation that explicitly models the 3D geometric relationships between various body parts using rotations and translations in 3D space. Since 3D rigid body motions are members of the special Euclidean group $SE(3)$, the proposed skeletal representation lies in the Lie group $SE(3)\times\ldots\times SE(3)$, which is a curved manifold. Using the proposed representation, human actions can be modeled as curves in this Lie group. Since classification of curves in this Lie group is not an easy task, we map the action curves from the Lie group to its Lie algebra, which is a vector space. We then perform classification using a combination of dynamic time warping, Fourier temporal pyramid representation and linear SVM. Experimental results on three action datasets show that the proposed representation performs better than many existing skeletal representations. The proposed approach also outperforms various state-of-the-art skeleton-based human action recognition approaches.

1. Introduction

Human action recognition has been an active area of research for the past several decades due to its applications in surveillance, video games, robotics, etc. In the past few decades, several approaches have been proposed for recognizing human actions from monocular RGB video sequences [1]. Unfortunately, the monocular RGB data is highly sensitive to various factors like illumination changes, variations in view-point, occlusions and background clutter. Moreover, monocular video sensors can not fully capture the human motion in 3D space. Hence, despite significant research efforts over the past few decades, action recognition still remains a challenging problem.

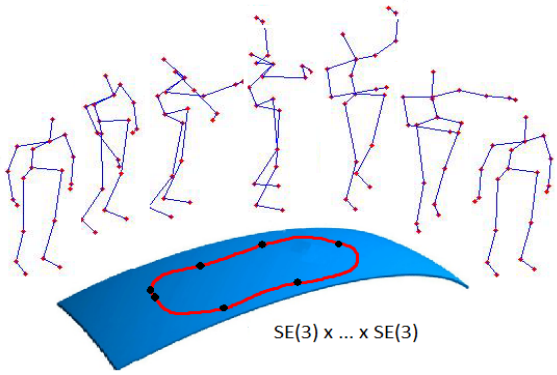

Figure 1: Representation of an action (skeletal sequence) as a curve in the Lie group $SE(3) \times \ldots \times SE(3)$.

A human body can be represented as an articulated system of rigid segments connected by joints, and human motion can be considered as a continuous evolution of the spatial configuration of these rigid segments [24]. Hence, if we can reliably extract and track the human skeleton, action recognition can be performed by classifying the temporal evolution of human skeleton. But, extracting the human skeleton reliably from monocular RGB videos is a very difficult task [9].

Sophisticated motion capture systems can be used to obtain the 3D locations of landmarks placed on the human body. But, such systems are very expensive, and require the user to wear a motion capture suit with markers which can hinder natural movements. With the recent advent of cost-effective depth sensors, extracting the human skeleton has become relatively easier. These sensors provide 3D depth data of the scene, which is robust to illumination changes and offers more useful information to recover 3D human skeletons. Recently, Shotton *et al*. [16] proposed a method to quickly and accurately estimate the 3D positions of skeletal joints using a single depth image. These recent advances have resulted in a renewed interest in skeleton-based human action recognition.

Existing skeleton-based action recognition approaches can be broadly grouped into two main categories: joint-based approaches and body part-based approaches. Inspired by the classical moving lights display experiment by Johansson [6], joint-based approaches consider the human skeleton simply as a set of points. These approaches try to model the motion of either individual joints or combinations of joints using various features like joint positions [5, 8], joint orientations with respect to a fixed coordinate axis [20], pairwise relative joint positions [19, 22], etc. On the other hand, body part-based approaches consider the human skeleton as a connected set of rigid segments (body parts). These approaches either model the temporal evolution of individual body parts [21] or focus on (directly) connected pairs of body parts and model the temporal evolution of joint angles [12, 13].

In this paper, we propose a new body part-based skeletal representation for action recognition. Inspired by the observation that for human actions, the relative geometry between various body parts (though not directly connected by a joint) provides a more meaningful description than their absolute locations (clapping is more intuitively described using the relative geometry between the two hands), we explicitly model the relative 3D geometry between different body parts in our skeletal representation. Given two rigid body parts, their relative geometry can be described using the rotation and translation required to take one body part to the position and orientation of the other (figure 3). Mathematically, rigid body rotations and translations in 3D space are members of the special Euclidean group $SE(3)$ [11], which is a matrix Lie group. Hence, we represent the relative geometry between a pair of body parts as a point in $SE(3)$, and the entire human skeleton as a point in the Lie group $SE(3) \times \ldots \times SE(3)$, where $\times$ denotes the direct product between Lie groups.

With the proposed skeletal representation, human actions can be modeled as curves (figure 1) in the Lie group $SE(3) \times \ldots \times SE(3)$, and action recognition can be performed by classifying these curves. Note that the Lie group $SE(3) \times \ldots \times SE(3)$ is a curved manifold and classification of curves in this space is not a trivial task. Moreover, standard classification approaches like SVM and temporal modeling approaches like Fourier analysis are not directly applicable to this curved space. To overcome these difficulties, we map the action curves from $SE(3) \times \ldots \times SE(3)$ to its Lie algebra $\mathfrak{se}(3) \times \ldots \times \mathfrak{se}(3)$, which is the tangent space at the identity element of the group.

Irrespective of the skeletal representation being used, classification of temporal sequences into different action categories is a difficult problem due to issues like rate variations, temporal misalignment, noise, etc. To handle rate variations, for each action category, we compute a nominal curve using dynamic time warping (DTW) [10], and warp all the curves to this nominal curve. To handle the temporal misalignment and noise issues, we represent the warped curves using the Fourier temporal pyramid (FTP) representation proposed in [19]. Final classification is performed using FTP and a linear SVM classifier. Figure 4 presents an overview of the proposed approach.

Contributions: 1) We represent human skeletons as points in the Lie group $SE(3) \times \ldots \times SE(3)$. The proposed representation explicitly models the 3D geometric relationships between various body parts using rotations and translations. 2) Since $SE(3) \times \ldots \times SE(3)$ is a curved manifold, we map all the action curves from the Lie group to its Lie algebra, and perform temporal modeling and classification in the Lie algebra. 3) We experimentally show that the proposed representation performs better than many existing skeletal representations by evaluating it on three different datasets: MSR-Action3D [7], UTKinect-Action dataset [20] and Florence3D-Action dataset [14]. We also show that the proposed approach outperforms various state-of-the-art skeleton-based human action recognition approaches.

Organization: We provide a brief review of the existing literature in section 2 and discuss the special Euclidean group $SE(3)$ in section 3. Section 4 presents the proposed skeletal representation and section 5 describes the temporal modeling and classification approach. We present our experimental results in section 6 and conclude the paper in section 7.

2. Relevant Work

In this section, we briefly review various skeleton-based human action recognition approaches. We refer the readers to [1] for a recent review of RGB video-based approaches and [23] for a recent review of depth map-based approaches.

Existing skeleton-based human action recognition approaches can be broadly grouped into two main categories: joint-based approaches and body part-based approaches. Joint-based approaches consider human skeleton as a set of points, whereas body part-based approaches consider human skeleton as a connected set of rigid segments. Approaches that use joint angles can be classified as part-based approaches since joint angles measure the geometry between (directly) connected pairs of body parts.

Joint-based approaches: Human skeletons were represented in [5] using the 3D joint locations, and the joint trajectories were modeled using a temporal hierarchy of covariance descriptors. A similar representation was used with Hidden Markov models (HMMs) in [8]. A set of 13 joint trajectories in a 4-D XYZT space was used in [15] to represent a human action, and their affine projections were compared using a subspace angles-based view-invariant similarity measure. In [19], a human skeleton was represented using pairwise relative positions of the joints, and the temporal evolutions of this representation were modeled using a

hierarchy of Fourier coefficients. Furthermore, an actionlet-based approach was used, in which discriminative joint combinations were selected using a multiple kernel learning approach. In [22], a human skeleton was represented using relative joint positions, temporal displacement of joints and offset of the joints with respect to the initial frame. Action classification was performed using the Naive-Bayes nearest neighbor rule in a lower dimensional space constructed using principal component analysis (PCA). A similar skeletal representation was used with random forests in [27]. A view invariant representation of human skeleton was obtained in [20] by quantizing the 3D joint locations into histograms based on their orientations with respect to a coordinate system fixed at the hip center. The temporal evolutions of this view-invariant representation were modeled using HMMs.

Part-based approaches: Human body was divided into five different parts in [21], and human actions were represented using the motion parameters of individual body parts like horizontal and vertical translations, in-plane rotations, etc. Principal component analysis was used to represent an action as a linear combination of a set of action basis, and classification was performed by comparing the PCA coefficients. In [2], a human skeleton was hierarchically divided into smaller parts and each part was represented using certain bio-inspired shape features. The temporal evolutions of these bio-inspired features were modeled using linear dynamical systems. Human skeleton was represented using 3D joint angles in [3], and the temporal evolutions of these angles were compared using DTW. In [12], few informative skeletal joints were automatically selected at each time instance based on highly interpretable measures such as mean or variance of the joint angles, maximum angular velocity of the joints, etc. Human actions were then represented as sequences of these informative joints, which were compared using the Levenshtein distance. Skeletal sequences were represented in [13] using pairwise affinities between joint angle trajectories, and then classified using linear SVM.

3. Special Euclidean Group $SE(3)$

In this section, we briefly discuss the special Euclidean group $SE(3)$. We refer the readers to [4] for a general introduction to Lie groups and [11] for further details on $SE(3)$ and rigid body kinematics.

The special Euclidean group, denoted by $SE(3)$, is the set of all 4 by 4 matrices of the form

$$P(R, \vec{d}) = \begin{bmatrix} R & \vec{d} \\ 0 & 1 \end{bmatrix}, \tag{1}$$

where $\vec{d} \in \mathcal{R}^3$, and $R \in \mathcal{R}^{3\times 3}$ is a rotation matrix. Members of $SE(3)$ act on points $z \in \mathcal{R}^3$ by rotating and translating them:

$$\begin{bmatrix} R & \vec{d} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} = \begin{bmatrix} Rz + \vec{d} \\ 1 \end{bmatrix}. \tag{2}$$

Elements of this set interact by the usual matrix multiplication, and from a geometrical point of view, can be smoothly organized to form a curved 6 dimensional manifold, giving them the structure of a Lie group [4]. The 4 by 4 identity matrix I_4 is a member of $SE(3)$ and is referred to as the identity element of this group.

The tangent plane to $SE(3)$ at the identity element I_4 is known as the Lie algebra of $SE(3)$, and is denoted by $\mathfrak{se}(3)$. It is a 6 dimensional vector space formed by all 4 by 4 matrices of the form $\begin{bmatrix} U & \vec{w} \\ 0 & 0 \end{bmatrix}$, where $\vec{w} \in \mathcal{R}^3$ and U is a 3 by 3 skew-symmetric matrix. For any element

$$B = \begin{bmatrix} U & \vec{w} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -u_3 & u_2 & w_1 \\ u_3 & 0 & -u_1 & w_2 \\ -u_2 & u_1 & 0 & w_3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \in \mathfrak{se}(3), \tag{3}$$

its vector representation $\text{vec}(B)$ is given by

$$\text{vec}(B) = [u_1, u_2, u_3, w_1, w_2, w_3]. \tag{4}$$

The exponential map $\exp_{SE(3)} : \mathfrak{se}(3) \rightarrow SE(3)$ and the logarithm map $\log_{SE(3)} : SE(3) \rightarrow \mathfrak{se}(3)$ between the Lie algebra $\mathfrak{se}(3)$ and the Lie group $SE(3)$ are given by

$$\begin{aligned} \exp_{SE(3)}(B) &= \mathbf{e}^B, \\ \log_{SE(3)}(P) &= \mathbf{log}(P), \end{aligned} \tag{5}$$

where $\mathbf{e}$ and $\mathbf{log}$ denote the usual matrix exponential and logarithm respectively. Since $\mathbf{log}(P)$ is not unique, we use the value with smallest norm. Please refer to [11] for efficient implementations of the exponential and logarithm maps of $SE(3)$.

Interpolation on $SE(3)$: Various approaches have been proposed in the past for interpolation on $SE(3)$ [25]. In this paper, we use a very simple piecewise interpolation scheme based on screw motions [26]. Given $Q_1, Q_2, \ldots, Q_n \in SE(3)$ at time instances $t_1, t_2, \ldots, t_n$ respectively, we use the following curve for interpolation:

$$\gamma(t) = Q_i \exp_{SE(3)}\left(\frac{t - t_i}{t_{i+1} - t_i} B_i\right) \text{ for } t \in [t_i, t_{i+1}], \tag{6}$$

where $B_i = \log_{SE(3)}\left(Q_i^{-1} Q_{i+1}\right)$ for $i = 1, 2, \ldots, n-1$.

$SE(3) \times \ldots \times SE(3)$: We can combine multiple $SE(3)$ using the direct product $\times$ to form a new Lie group $\mathcal{M} = SE(3) \times \ldots \times SE(3)$ with identity element $(I_4, \ldots, I_4)$ and Lie algebra $\mathfrak{m} = \mathfrak{se}(3) \times \ldots \times \mathfrak{se}(3)$. The exponential and logarithm maps for $(B_1, \ldots, B_K) \in \mathfrak{m}$ and $(P_1, \ldots, P_K) \in \mathcal{M}$ are given by

$$\begin{aligned} &\exp_{\mathcal{M}}((B_1, \ldots, B_K)) = (\mathbf{e}^{B_1}, \ldots, \mathbf{e}^{B_K}), \\ &\log_{\mathcal{M}}((P_1, \ldots, P_K)) = (\mathbf{log}(P_1), \ldots, \mathbf{log}(P_K)). \end{aligned} \tag{7}$$

Interpolation on $SE(3) \times \ldots \times SE(3)$ can be performed by simultaneously interpolating on individual $SE(3)$.

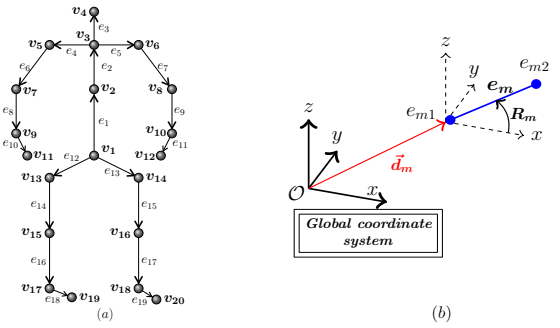

Figure 2: (a) An example skeleton consisting of 20 joints and 19 body parts, (b) Representation of a body part e_m in the global coordinate system.

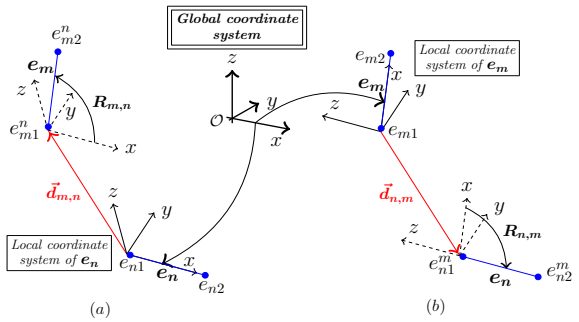

Figure 3: (a) Representation of body part e_m in the local coordinate system of e_n, (b) Representation of body part e_n in the local coordinate system of e_m.

4. Proposed Skeletal Representation

Let $S = (V, E)$ be a skeleton, where $V = \{v_1, \ldots, v_N\}$ denotes the set of joints and $E = \{e_1, \ldots, e_M\}$ denotes the set of oriented rigid body parts. Figure 2 shows an example skeleton with 20 joints and 19 body parts. Let $e_{n1} \in \mathcal{R}^3$, $e_{n2} \in \mathcal{R}^3$ respectively denote the starting and end points of body part e_n. Let l_n denote the length of e_n.

Given a pair of body parts e_m and e_n, to describe their relative geometry, we represent each of them in a local coordinate system attached to the other. Figure 3 explains this pictorially. The local coordinate system of body part e_n is obtained by rotating (with minimum rotation) and translating the global coordinate system such that e_{n1} becomes the origin and e_n coincides with the x-axis (refer to figure 3(a)). Let $e^n_{m1}(t), e^n_{m2}(t) \in \mathcal{R}^3$ respectively denote the starting and end points of e_m represented in the local coordinate system attached to e_n at time instance t. Then

$$\begin{bmatrix} e^n_{m1}(t) & e^n_{m2}(t) \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} R_{m,n}(t) & \vec{d}_{m,n}(t) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & l_m \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}, \quad (8)$$

where $R_{m,n}(t)$ and $\vec{d}_{m,n}(t)$ are the rotation and translation (measured in the local coordinate system attached to e_n) required to take e_n to the position and orientation of e_m.

Similarly, we can represent e_n in the local coordinate system attached to e_m (refer to figure 3(b)). Let $e^m_{n1}(t), e^m_{n2}(t) \in \mathcal{R}^3$ respectively denote the starting and end points of e_n represented in the local coordinate system attached to e_m at time instance t. Then

$$\begin{bmatrix} e^m_{n1}(t) & e^m_{n2}(t) \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} R_{n,m}(t) & \vec{d}_{n,m}(t) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & l_n \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}, \quad (9)$$

where $R_{n,m}(t)$ and $\vec{d}_{n,m}(t)$ are the rotation and translation (measured in the local coordinate system attached to e_m) required to take e_m to the position and orientation of e_n.

Since the lengths of body parts do not vary with time, the relative geometry between e_m and e_n at time instance t can be described using

$$\begin{aligned} P_{m,n}(t) &= \begin{bmatrix} R_{m,n}(t) & \vec{d}_{m,n}(t) \\ 0 & 1 \end{bmatrix} \in SE(3), \\ P_{n,m}(t) &= \begin{bmatrix} R_{n,m}(t) & \vec{d}_{n,m}(t) \\ 0 & 1 \end{bmatrix} \in SE(3). \end{aligned} \quad (10)$$

At first glance it might appear that using only $P_{m,n}(t)$ or $P_{n,m}(t)$ would suffice. Consider the case in which e_n is rotating about an axis along e_m. Though there is relative motion between the two, the matrix $P_{m,n}(t)$ will not change. Similarly, if e_m is rotating about an axis along e_n, then the matrix $P_{n,m}(t)$ will not change. So, if we represent the relative geometry using only $P_{m,n}(t)$ or $P_{n,m}(t)$, the representation will not change under certain kinds of relative motions, which is undesirable. Hence, we use both $P_{m,n}(t)$ and $P_{n,m}(t)$ to represent the relative geometry between e_m and e_n. Note that both $P_{m,n}(t)$ and $P_{n,m}(t)$ do not change only when both e_m and e_n undergo same rotation and translation, i.e., only when there is no relative motion between them.

Using the relative geometry between all pairs of body parts, we represent a skeleton S at time instance t using $C(t) = (P_{1,2}(t),\ P_{2,1}(t) \ldots, P_{M-1,M}(t), P_{M,M-1}(t)) \in SE(3) \times \ldots \times SE(3)$, where M is the number of body parts. Using the proposed skeletal representation, a skeletal sequence describing an action can be represented (figure 1) as a curve $\{C(t),\ t \in [0, T]\}$ in $SE(3) \times \ldots \times SE(3)$. Classification of action curves in the curved space $SE(3) \times \ldots \times SE(3)$ is not a trivial task. Moreover, standard classification approaches like SVM and temporal modeling approaches like Fourier analysis are not directly applicable to this space. To overcome these difficulties, we map the action curves from $SE(3) \times \ldots \times E(3)$ to its Lie algebra $\mathfrak{se}(3) \times \ldots \times \mathfrak{se}(3)$, which is the tangent space at the identity element. The Lie algebra curve (in vector representation) corresponding to $C(t)$ is given by

$$\begin{aligned} \mathfrak{C}(t) = \big[&\text{vec}(\mathbf{log}(P_{1,2}(t))), \text{vec}(\mathbf{log}(P_{2,1}(t))), \ldots, \\ &\text{vec}(\mathbf{log}(P_{M-1,M}(t))), \text{vec}(\mathbf{log}(P_{M,M-1}(t))) \big]. \end{aligned} \quad (11)$$

At any time instance t, $\mathfrak{C}(t)$ is a vector of dimension $6M(M-1)$. Hence, we represent actions as temporal evolutions of $6M(M-1)$-dimensional vector.

Note that, we are using only the relative measurements $P_{m,n}(t)$ in our skeletal representation. We also performed experiments by adding the absolute 3D locations of body parts to the skeletal representation. The 3D location of a rigid body part e_m can be described using its rotation R_m with respect to global x-axis and the translation $\vec{d}_m$ of its starting point e_{m1} from the origin (refer to figure 2(b)). But, using the absolute locations of body parts did not give any improvement, suggesting that the information about absolute locations is redundant for the actions used in our experiments. Hence, we just use the relative measurements in this paper.

5. Temporal Modeling and Classification

Classification of curves in the Lie algebra into different action categories is not straightforward due to various issues like rate variations, temporal misalignment, noise, etc. Following [17], we use DTW [10] to handle rate variations. During training, for each action category, we compute a nominal curve using the algorithm described in Table 1, and warp all the training curves to this nominal curve using DTW. We use the squared Euclidean distance in the Lie algebra for DTW. Note that to compute a nominal curve all the curves should have equal number of samples. For this, we use the interpolation algorithm presented in section 3 and re-sample the curves in $SE(3) \times \ldots \times SE(3)$ before mapping them to Lie algebra. To handle the temporal misalignment and noise issues, we represent the warped curves using the recently proposed Fourier temporal pyramid representation [19] removing the high frequency coefficients. We apply FTP for each dimension separately and concatenate all the Fourier coefficients to obtain the final feature vector. Action recognition is performed by classifying the final feature vectors using one-vs-all linear SVM. Figure 4 gives an overview of the entire approach.

6. Experimental Evaluation

In this section, we evaluate the proposed skeletal representation using three different datasets: MSR-Action3D [7], UTKinect-Action [20] and Florence3D-Action [14]. The code used for our experiments can be downloaded from http://ravitejav.weebly.com/kbac.html.

MSR-Action3D dataset [7]: This dataset was captured using a depth sensor similar to Kinect. It consists of 20 actions performed by 10 different subjects. Each subject performed every action two or three times. Altogether, there are 557 action sequences. The 3D locations of 20 joints are provided with the dataset. This is a challenging dataset because many of the actions are highly similar to each other.

Table 1: Algorithm for computing a nominal curve

Input: Curves $\mathfrak{C}_1(t), \ldots, \mathfrak{C}_J(t)$ at $t = 0, 1, \ldots, T$. Maximum number of iterations max and threshold δ.

Output: Nominal curve $\mathfrak{C}(t)$ at $t = 0, 1, \ldots, T$.

Initialization: $\mathfrak{C}(t) = \mathfrak{C}_1(t)$, iter = 0.
while iter $< max$
  Warp each curve $\mathfrak{C}_j(t)$ to the nominal curve $\mathfrak{C}(t)$ using DTW with squared Euclidean distance to get a warped curve $\mathfrak{C}_j^w(t)$.
  Compute a new nominal $\mathfrak{C}^{'}(t)$ using
  $\mathfrak{C}^{'}(t) = \frac{1}{J}\sum_{j=1}^{J} \mathfrak{C}_j^w(t)$.
  if $\sum_{t=0}^{T} \|\mathfrak{C}^{'}(t) - \mathfrak{C}(t)\|_2^2 \leq \delta$ ($\|\cdot\|_2$ denotes ℓ_2 norm)
  break
  end
  $\mathfrak{C}(t) = \mathfrak{C}^{'}(t)$; iter = iter + 1;
end

UTKinect-Action dataset [20]: This dataset was captured using a stationary Kinect sensor. It consists of 10 actions performed by 10 different subjects. Each subject performed every action twice. Altogether, there are 199 action sequences. The 3D locations of 20 joints are provided with the dataset. This is a challenging dataset due to variations in the view point and high intra-class variations.

Florence3D-Action dataset [14] This dataset was captured using a stationary Kinect sensor. It consists of 9 actions performed by 10 different subjects. Each subject performed every action two or three times. Altogether, there are 215 action sequences. The 3D locations of 15 joints are provided with the dataset. This is a challenging dataset due to high intra-class variations (same action is performed using left hand in some sequences and right hand in some other) and the presence of actions like *drink from a bottle* and *answer phone* which are quite similar to each other.

Basic pre-processing: In the case of MSR-Action3D and UTKinect-Action datasets, each skeleton has 19 parts and 20 joints, whereas in the case of Florence3D-Action dataset, each skeleton has 14 parts and 15 joints. To make the skeletal data invariant to absolute location of the human in the scene, all 3D joint coordinates were transformed from the world coordinate system to a person-centric coordinate system by placing the hip center at the origin. For each dataset, we took one of the skeletons as reference, and normalized all the other skeletons (without changing their joint angles) such that their body part lengths are equal to the corresponding lengths of the reference skeleton. This normalization makes the skeletons scale-invariant. We also rotated the skeletons such that the ground plane projection of the vector from left hip to right hip is parallel to the global x-axis. This rotation makes the skeletons view-invariant.

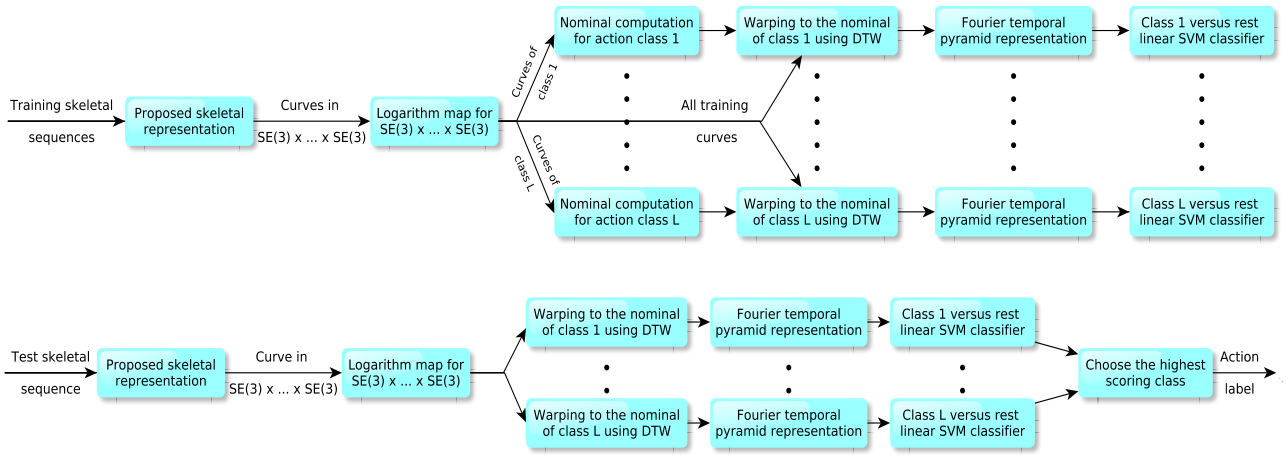

Figure 4: The top row shows all the steps involved in training and the bottom row shows all the steps involved in testing.

6.1. Alternative Skeletal Representations

To show the effectiveness of the proposed skeletal representation, we compare it with the following four alternative skeletal representations:

Joint positions (JP): Concatenation of 3D coordinates of all the joints $v_1, \ldots, v_N$.

Pairwise relative positions of the joints (RJP): Concatenation of all the vectors $\overrightarrow{v_i v_j}$, $1 \leq i < j \leq N$.

Joint angles (JA): Concatenation of the quaternions corresponding to all joint angles. We also tried Euler angles and Euler axis-angle representations for the joint angles, but quaternions gave the best results.

Individual body part locations (BPL): Each individual body part is represented as a point in $SE(3)$ using its rotation and translation relative to the global x-axis. Mapping the points from $SE(3)$ to $\mathfrak{se}(3)$, we get a 6M-dimensional vector representation, where M is the number of body parts.

For fair comparison, we used the temporal modeling and classification approach described in section 5 with all the representations.

6.2. Evaluation Settings and Parameters

For MSR-Action3D dataset, we followed the cross-subject test setting of [7], in which half of the subjects were used for training and the other half were used for testing. Following [7], we divided the dataset into subsets AS_1, AS_2 and AS_3, each consisting of 8 actions, and performed recognition on each subset separately. The subsets AS_1 and AS_2 were intended to group actions with similar movements, while the subset AS_3 was intended to group complex actions together.

For UTKinect-Action and Florence3D-Action datasets, we followed the cross-subject test setting of [27], in which half of the subjects were used for training and the remaining half were used for testing.

In all the experiments, we used a three-level Fourier temporal pyramid with 1/4 length of each segment as low-frequency coefficients. The value of SVM parameter C was set to 1 in all the experiments. As explained in section 5, for each dataset, all the curves in $SE(3) \times \ldots \times SE(3)$ were re-sampled to have same length. The reference length was chosen to be the maximum number of samples in any curve in the dataset before re-sampling. All the results reported in this paper were averaged over ten different combinations of training and test data.

6.3. Results

Comparison with other skeletal representations: Table 2 reports the recognition rates for various skeletal representations on MSR-Action3D dataset. The recognition rates in the last row are the average of the recognition rates for the three subsets AS_1, AS_2 and AS_3. We can see that the proposed representation performs better than joint positions, joint angles and individual body part locations, and slightly worse (0.46%) when compared to relative joint positions. The average accuracy of the proposed representation is 10.61% better than the average accuracy of joint angles, 3.65% better than the average accuracy of individual body part locations, and 0.71% better than the average accuracy of joint positions.

Table 3 reports the recognition rates for various skeletal representations on UTKinect-Action and Florence3D-Action datasets. In the case of UTKinect-Action dataset, the average accuracy of the proposed representation is 3% better than the average accuracy of joint angles, 2.5% better than the average accuracy of individual body part locations, 2.4% better than the average accuracy of joint positions, and 1.5% better that the average accuracy of relative joint positions. In the case of Florence3D-Action dataset, the average accuracy of the proposed representation is 9.5% better than the average accuracy of joint angles, 10.1% better

Table 2: Recognition rates for various skeletal representations on MSR-Action3D dataset using the protocol of [7]

Dataset	JP	RJP	JA	BPL	Proposed
AS_1	93.36	95.77	84.51	90.30	94.72
AS_2	85.53	86.90	68.05	83.91	86.83
AS_3	99.55	99.28	96.17	95.39	99.02
Average	92.81	**93.98**	82.91	89.87	93.52

Table 3: Recognition rates for various skeletal representations on UTKinect-Action and Florence3D-Action datasets

Dataset	JP	RJP	JA	BPL	Proposed
UTKinect	94.68	95.58	94.07	94.57	**97.08**
Florence3D	85.26	85.20	81.36	80.80	**90.88**

than the average accuracy of individual body part locations, 5.6% better than the average accuracy of joint positions, and 5.7% better that the average accuracy of relative joint positions. These results clearly demonstrate the superiority of the proposed representation over various existing skeletal representations.

Figure 5 shows the confusion matrices for MSR-Action3D AS_1, MSR-Action3D AS_2 and Florence3D Action datasets. We skip MSR-Action3D AS_3 and UTKinect-Action datasets as the corresponding recognition rates are very high. We can see that most of the confusions are between highly similar actions like *hammer and high throw* in the case of MSR-Action3D AS_1, *draw X, draw tick, draw circle, hand catch and side boxing* in the case of MSR-Action3D AS_2, and *drink, answer phone and read watch* in the case of Florence3D-Action dataset.

Comparison with state-of-the-art results:
Table 4 compares the proposed approach with various state-of-the-art skeleton-based human action recognition approaches. We can see that the proposed approach gives the best results on all datasets. Specifically, it outperforms the state-of-the-art by 6.1% on UTKinect-Action dataset and by 8.8% on Florence3D-Action dataset.

Note that we have reported two different recognition rates for the proposed approach on MSR-Action3D dataset. The recognition rate of 93.52% corresponds to the experimental setting of [7] and the recognition rate of 89.48% corresponds to the experimental setting of [19]. In [19], instead of dividing the dataset into three subsets, the actionlet-based approach was applied to the entire dataset consisting of 20 actions. This experimental setting is more difficult compared to that of [7].

Some recent approaches like [13, 27] have reported recognition rates around 94.5% for MSR-Action3D dataset by combining skeletal features with additional depth-based features. Since this paper's focus is not on combining multiple features, we only use the skeleton-based results reported in [13, 27] for comparison.

It is interesting to note that even joint positions and relative joint positions (when used with the temporal modeling and classification approach presented in section 5) produce results better than the state-of-the-art reported on UTKinect-Action and Florence3D-Action datasets. This suggests that the combination of DTW, FTP and linear SVM is well-suited for skeleton-based action classification.

Table 4: Comparison with the state-of-the-art results

MSR-Action3D dataset (protocol of [7])	
Histograms of 3D joints [20]	78.97
EigenJoints [22]	82.30
Joint angle similarities [13]	83.53
Spatial and temporal part-sets[18]	90.22
Covariance descriptors [5]	90.53
Random forests [27]	90.90
Proposed approach	**93.52**
MSR-Action3D dataset (protocol of [19])	
Actionlets [19]	88.20
Proposed approach	**89.48**
UTKinect-Action dataset	
Histograms of 3D joints [20]	90.92
Random forests [27]	87.90
Proposed approach	**97.08**
Florence3D-Action dataset	
Multi-Part Bag-of-Poses [14]	82.00
Proposed approach	**90.88**

7. Conclusion and Future Work

In this paper, we represented a human skeleton as a point in the Lie group $SE(3) \times \ldots \times SE(3)$, by explicitly modeling the 3D geometric relationships between various body parts using rotations and translations. Using the proposed skeletal representation, we modeled human actions as curves in this Lie group. Since $SE(3) \times \ldots \times SE(3)$ is a curved manifold, we mapped all the curves to its Lie algebra, which is a vector space, and performed temporal modeling and classification in the Lie algebra. We experimentally showed that the proposed representation performs better than many existing skeletal representations on three different action datasets. We also showed that the proposed approach outperforms various state-of-the-art skeleton-based human action recognition approaches.

In our work, we used the relative geometry between all pairs of body parts. But, each action is usually characterized by the interactions of a specific set of body parts. Hence, we are planning to explore various strategies to automatically identify the set of body parts that differentiates a given action from the rest. In this paper, we focused only on actions performed by a single person. We are planning to extend this representation to model multi-person interactions.

Acknowledgements: This research was supported by a MURI from the US Office of Naval Research under the grant 1141221258513.

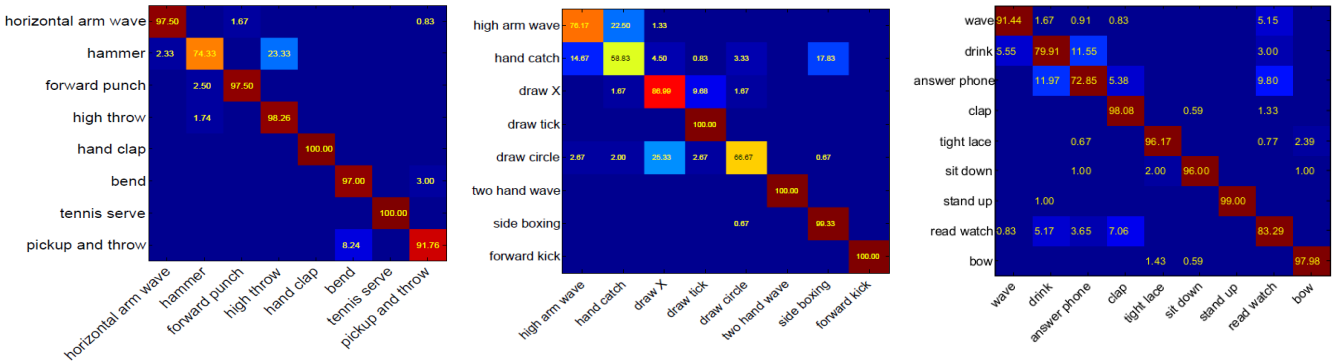

Figure 5: Confusion matrices: Left – MSR-Action3D AS_1; Center - MSR-Action3D AS_2; Right – Florence3D-Action

References

[1] J. K. Aggarwal and M. S. Ryoo. Human Activity Analysis: A Review. *ACM Computing Surveys*, 43(3):16:1–16:43, 2011. 2, 3

[2] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal. Bio-inspired Dynamic 3D Discriminative Skeletal Features for Human Action Recognition. In *CVPRW*, 2013. 4

[3] D. M. Gavrila and L. S. Davis. Towards 3-D Model-based Tracking and Recognition of Human Movement: A Multi-view Approach. In *International Workshop on Automatic Face and Gesture Recognition*, 1995. 4

[4] B. Hall. Lie Groups, Lie Algebras, and Representations: An Elementary Introduction. *Springer*, 2003. 4

[5] M. Hussein, M. Torki, M. Gowayyed, and M. El-Saban. Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations. In *IJCAI*, 2013. 3, 8

[6] G. Johansson. Visual Perception of Biological Motion and a Model for its Analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. 3

[7] W. Li, Z. Zhang, and Z. Liu. Action Recognition Based on a Bag of 3D Points. In *CVPRW*, 2010. 1, 3, 6, 7, 8

[8] F. Lv and R. Nevatia. Recognition and Segmentation of 3D Human Action Using HMM and Multi-class Adaboost. In *ECCV*, 2006. 3

[9] T. B. Moeslund, A. Hilton, and V. Krüger. A Survey of Advances in Vision-based Human Motion Capture and Analysis. *CVIU*, 104(2-3):90–126, 2006. 2

[10] M. Müller. Information Retrieval for Music and Motion. *Springer-Verlag New York, Inc.*, 2007. 3, 6

[11] R. M. Murray, Z. Li, and S. S. Sastry. A Mathematical Introduction to Robotic Manipulation. *CRC press*, 1994. 3, 4

[12] 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 *CVPRW*, 2012. 3, 4

[13] E. Ohn-bar and M. M. Trivedi. Joint Angles Similarities and HOG^2 for Action Recognition. In *In CVPRW*, 2013. 3, 4, 8

[14] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala. Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses. In *CVPRW*, 2013. 3, 6, 8

[15] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the Space of a Human Action. In *ICCV*, 2005. 3

[16] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time Human Pose Recognition in Parts From a Single Depth Image. In *CVPR*, 2011. 2

[17] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, and R. Chellappa. Rate-invariant Recognition of Humans and Their Activities. *IEEE Trans. on Image Processing*, 18(6):1326–1339, 2009. 6

[18] C. Wang, Y. Wang, and A. L. Yuille. An Approach to Pose-based Action Recognition. In *CVPR*, 2013. 8

[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *CVPR*, 2012. 1, 3, 6, 8

[20] L. Xia, C. C. Chen, and J. K. Aggarwal. View Invariant Human Action Recognition Using Histograms of 3D Joints. In *CVPRW*, 2012. 3, 4, 6, 8

[21] Y. Yacoob and M. J. Black. Parameterized Modeling and Recognition of Activities. In *ICCV*, 1998. 3, 4

[22] X. Yang and Y. Tian. EigenJoints-based Action Recognition Using Naïve-Bayes-Nearest-Neighbor. In *CVPRW*, 2012. 3, 4, 8

[23] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. A Survey on Human Motion Analysis from Depth Data. In *CVPR Tutorial on RGBD Cameras*, 2013. 3

[24] V. M. Zatsiorsky. Kinematics of Human Motion. *Human Kinetics Publishers*, 1997. 2

[25] M. Zefran and V. Kumar. Two Methods for Interpolating Rigid Body Motions. In *ICRA*, 1998. 4

[26] M. Zefran, V. Kumar, and C. Croke. Choice of Riemannian Metrics for Rigid Body Kinematics. In *ASME Design Engineering Technical Conference and Computers in Engineering Conference*, 1996. 4

[27] Y. Zhu, W. Chen, and G. Guo. Fusing Spatiotemporal Features and Joints for 3D Action Recognition. In *CVPRW*, 2013. 4, 7, 8