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

Raviteja Vemulapalli
University of Maryland, College Park.

Professor Rama Chellappa

Dr. Felipe Arrate
Action Recognition from 3D Skeletal Data

- **Motivation:** Humans can recognize many actions directly from skeletal sequences.

But, how do we get the 3D skeletal data?
Cost Effective Depth Sensors

- Human performing an action
- Cost effective depth sensors like Kinect
- State-of-the-art depth-based skeleton estimation algorithm [Shotton 2011]
- Real-time skeletal sequence


UTKinect-Action dataset [Xia2012]
Applications

Gesture-based Control

Elderly Care

Teaching Robots
Overview of a typical skeleton-based action recognition approach.
How to represent a 3D human skeleton for action recognition?
Human Skeleton: Points or Rigid Rods?

Set of points (joints)

Set of rigid rods (body parts)
Human Skeleton as a Set of Points

- Inspired by the moving lights display experiment by [Johansson 1973].
- Popularly-used skeletal representation.

Representation:
Concatenation of the 3D coordinates of the joints.

Human skeleton is a set of 3D rigid rods (body parts) connected by joints.

Spatial configuration of these rods can be represented using joint angles (shown using red arcs in the below figure).

**Representation:**
Concatenation of the Euler angles or Axis-angle or quaternion representations corresponding to the 3D joint angles.
Human actions are characterized by how different body parts move relative to each other.

For action recognition, we need a skeletal representation whose temporal evolution directly describes the relative motion between various body parts.
Proposed Skeletal Representation

We represent a skeleton using the relative 3D geometry between different body parts.

The relative geometry between two body parts can be described using the 3D rotation and translation required to take one body part to the position and orientation of the other.
We describe the relative geometry between two rigid body parts \((e_m, e_n)\) at time instance \(t\) using the rotation \(R_{m,n}(t)\) and the translation \(\vec{d}_{m,n}(t)\) (measured in the local coordinate system attached to \(e_n\)) required to take \(e_n\) to the position and orientation of \(e_m\).

\[
\begin{bmatrix}
    e^m_{m1}(t) & e^m_{m2}(t) \\
    1 & 1
\end{bmatrix} =
\begin{bmatrix}
    R_{m,n}(t) & \vec{d}_{m,n}(t) \\
    0 & 1
\end{bmatrix}
\begin{bmatrix}
    e^n_{n1}(t) & e^n_{n2}(t) \\
    1 & 1
\end{bmatrix}
\]

Rotation and translation vary with time.

Scaling factor: Independent of time since lengths of the body parts do not change with time.
Rigid body rotations and translations are members of special Euclidean group SE(3)
The special Euclidean group, denoted by $SE(3)$, is the set of all $4 \times 4$ matrices of the form

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

where $\hat{d} \in \mathbb{R}^3$ and $R$ is a $3 \times 3$ rotation matrix.

The group $SE(3)$ is a smooth 6-dimensional curved manifold.

The tangent plane to the manifold $SE(3)$ at the identity matrix $I_4$, denoted by $\mathfrak{se}(3)$, is known as the Lie algebra of $SE(3)$.

Lie algebra $\mathfrak{se}(3)$ is a 6-dimensional vector space.

The exponential map $exp_{SE(3)}: \mathfrak{se}(3) \to SE(3)$ and the logarithm map $log_{SE(3)}: SE(3) \to \mathfrak{se}(3)$ are given by

$$exp_{SE(3)}(B) = e^B,$$
$$log_{SE(3)}(P) = \log(P),$$

where $e$ and $\log$ denote the usual matrix exponential and logarithm.
Human skeleton is described using the relative 3D geometry between all pairs of body parts.

\[
\{ p_{m,n}(t) = \begin{bmatrix} R_{m,n}(t) & \vec{d}_{m,n}(t) \\ 0 & 1 \end{bmatrix} : m \neq n, 1 \leq m, n \leq 19 \} \in SE(3) \times \cdots \times SE(3)
\]

Point in $SE(3)$ describing the relative 3D geometry between body parts $(e_m, e_n)$ at time instance $t$. Lie group obtained by combining multiple $SE(3)$ using the direct product $\times$. 

Proposed Skeletal Representation
Using the proposed skeletal representation, a skeletal sequence can be represented as a curve in the Lie group $SE(3) \times \cdots \times SE(3)$:

$$\{ \{ P_{m,n}(t) | m \neq n, 1 \leq m, n \leq 19 \} | t \in [0, T] \}.$$
Proposed Action Representation

- Classification of the curves in $SE(3) \times \cdots \times SE(3)$ into different action categories is a difficult task due to the non-Euclidean nature of the space.

- Standard classification approaches like support vector machines (SVM) and temporal modeling approaches like Fourier analysis are not directly applicable to this space.

- To overcome these difficulties, we map the curves from the Lie group $SE(3) \times \cdots \times SE(3)$ to its Lie algebra $\mathfrak{se}(3) \times \cdots \times \mathfrak{se}(3)$, which is a vector space.
Proposed Action Representation

- Human actions are represented as curves in the Lie algebra $\mathfrak{se}(3) \times \cdots \times \mathfrak{se}(3)$.

$$\{ \log_{SE(3)}(P_{m,n}(t)) \mid m \neq n, 1 \leq m, n \leq 19 \}, t \in [0, T].$$

- Action recognition can be performed by classifying the curves in the vector space $\mathfrak{se}(3) \times \cdots \times \mathfrak{se}(3)$ into different action categories.
Temporal Modeling and Classification

- Action classification is a difficult task due to various issues like rate variations, temporal misalignments, noise, etc.

- Following [Veeraraghavan 2009], we use Dynamic Time Warping (DTW) to handle rate variations.

- Following [Wang 2012], we use the Fourier temporal pyramid (FTP) representation to handle noise and temporal misalignments.

- We use linear SVM with Fourier temporal pyramid representation for final classification.


We interpolate all the curves in the Lie group $SE(3) \times \cdots \times SE(3)$ to have same length.

\begin{center}
\begin{tabular}{|l|}
\hline
\textbf{Input:} Curves $\mathcal{C}_1(t), \ldots, \mathcal{C}_J(t)$ at $t = 0, 1, \ldots, T$. \\
Maximum number of iterations $max$ and threshold $\delta$. \\
\hline
\textbf{Output:} Nominal curve $\mathcal{C}(t)$ at $t = 0, 1, \ldots, T$. \\
\hline
\textbf{Initialization:} $\mathcal{C}(t) = \mathcal{C}_1(t)$, $\text{iter} = 0$. \\
\textbf{while} $\text{iter} < max$ \\
\hspace{1em} Warp each curve $\mathcal{C}_j(t)$ to the nominal curve $\mathcal{C}(t)$ using DTW with squared Euclidean distance to get a warped curve $\mathcal{C}^w_j(t)$. \\
\hspace{1em} Compute a new nominal $\mathcal{C}'(t)$ using \\
\hspace{1em} $\mathcal{C}'(t) = \frac{1}{J} \sum_{j=1}^{J} \mathcal{C}^w_j(t)$. \\
\hspace{1em} \textbf{if} $\sum_{t=0}^{T} ||\mathcal{C}'(t) - \mathcal{C}(t)||^2 \leq \delta$ (\text{$\ell_2$ norm}) \\
\hspace{1em} \textbf{break} \\
\hspace{1em} \textbf{end} \\
\hspace{1em} $\mathcal{C}(t) = \mathcal{C}'(t)$; $\text{iter} = \text{iter} + 1$; \\
\textbf{end} \\
\hline
\end{tabular}
\end{center}
Fourier Temporal Pyramid Representation

Magnitude of the low frequency Fourier coefficients from each level are used to represent a time sequence.

Overview of the Proposed Approach

Training skeletal sequences → Proposed skeletal representation → Curves in SE(3) x ... x SE(3) → Logarithm map for SE(3) x ... x SE(3)

Nominal computation for action class 1 → Warping to the nominal of class 1 using DTW → Fourier temporal pyramid representation → Class 1 versus rest linear SVM classifier

Nominal computation for action class L → Warping to the nominal of class L using DTW → Fourier temporal pyramid representation → Class L versus rest linear SVM classifier

Test skeletal sequence → Proposed skeletal representation → Curve in SE(3) x ... x SE(3) → Logarithm map for SE(3) x ... x SE(3)

Warping to the nominal of class 1 using DTW → Fourier temporal pyramid representation → Class 1 versus rest linear SVM classifier

Warping to the nominal of class L using DTW → Fourier temporal pyramid representation → Class L versus rest linear SVM classifier

Choose the highest scoring class → Action label
Experiments: Datasets

MSR-Action3D dataset
- Total 557 action sequences
- 20 actions
- 10 subjects


UTKinect-Action dataset
- Total 199 action sequences
- 10 actions
- 10 subjects


Florence3D-Action dataset
- Total 215 action sequences
- 9 actions
- 10 subjects

Joint positions (JP):
Concatenation of the 3D coordinates of the joints.

Pairwise relative positions of the joints (RJP):
Concatenation of the 3D vectors $\vec{v}_i \vec{v}_j, 1 \leq i < j \leq 20$.

Joint angles (JA):
Concatenation of the quaternions corresponding to the joint angles (shown using red arcs in the figure).

Individual body part locations (BPL):
Each body part $e_m$ is represented as a point in $SE(3)$ using its relative 3D geometry with respect to the global $x$-axis.
MSR-Action3D Dataset

- Total 557 action sequences: 20 actions performed (2 or 3 times) by 10 different subjects.
- Dataset is further divided into 3 subsets: AS1, AS2 and AS3.

<table>
<thead>
<tr>
<th>Action Set 1 (AS1)</th>
<th>Action Set 2 (AS2)</th>
<th>Action Set 3 (AS3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal arm wave</td>
<td>High arm wave</td>
<td>High throw</td>
</tr>
<tr>
<td>Hammer</td>
<td>Hand catch</td>
<td>Forward kick</td>
</tr>
<tr>
<td>Forward punch</td>
<td>Draw x</td>
<td>Side kick</td>
</tr>
<tr>
<td>High throw</td>
<td>Draw tick</td>
<td>Jogging</td>
</tr>
<tr>
<td>Hand clap</td>
<td>Draw circle</td>
<td>Tennis swing</td>
</tr>
<tr>
<td>Bend</td>
<td>Two hand wave</td>
<td>Tennis serve</td>
</tr>
<tr>
<td>Tennis serve</td>
<td>Forward wave</td>
<td>Golf swing</td>
</tr>
<tr>
<td>Pickup &amp; throw</td>
<td>Forward kick</td>
<td>Pickup &amp; throw</td>
</tr>
</tbody>
</table>
Results: MSR-Action3D Dataset

- Experiments performed on each of the subsets (AS1, AS2 and AS3) separately.
- Half of the subjects were used for training and the other half were used for testing.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>JP</th>
<th>RJP</th>
<th>JA</th>
<th>BPL</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS1</td>
<td>91.65</td>
<td>92.15</td>
<td>85.80</td>
<td>83.87</td>
<td><strong>95.29</strong></td>
</tr>
<tr>
<td>AS2</td>
<td>75.36</td>
<td>79.24</td>
<td>65.47</td>
<td>75.23</td>
<td><strong>83.87</strong></td>
</tr>
<tr>
<td>AS3</td>
<td>94.64</td>
<td>93.31</td>
<td>94.22</td>
<td>91.54</td>
<td><strong>98.22</strong></td>
</tr>
<tr>
<td>Average</td>
<td>87.22</td>
<td>88.23</td>
<td>81.83</td>
<td>83.54</td>
<td><strong>92.46</strong></td>
</tr>
</tbody>
</table>

Recognition rates for various skeletal representations on MSR-Action3D dataset.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigen Joints</td>
<td>82.30</td>
</tr>
<tr>
<td>Joint angle similarities</td>
<td>83.53</td>
</tr>
<tr>
<td>Spatial and temporal part-sets</td>
<td>90.22</td>
</tr>
<tr>
<td>Covariance descriptors on 3D joint locations</td>
<td>90.53</td>
</tr>
<tr>
<td>Random forests</td>
<td>90.90</td>
</tr>
<tr>
<td><strong>Proposed approach</strong></td>
<td><strong>92.46</strong></td>
</tr>
</tbody>
</table>

Comparison with the state-of-the-art results on MSR-Action3D dataset.
MSR-Action3D Confusion Matrices

Action set 1 (AS1)
Average recognition accuracy: 95.29%

Action set 2 (AS2)
Average recognition accuracy: 83.87%

Action set 3 (AS3)
Average recognition accuracy: 98.22%
Results: UTKinect-Action Dataset

- Total 199 action sequences: 10 actions performed (2 times) by 10 different subjects.
- Half of the subjects were used for training and the other half were used for testing.

<table>
<thead>
<tr>
<th>JP</th>
<th>RJP</th>
<th>JA</th>
<th>BPL</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>94.68</td>
<td>95.58</td>
<td>94.07</td>
<td>94.57</td>
<td>97.08</td>
</tr>
</tbody>
</table>

Recognition rates for various skeletal representations on UTKinect-Action dataset.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forests</td>
<td>87.90</td>
</tr>
<tr>
<td>Histograms of 3D joints</td>
<td>90.92</td>
</tr>
<tr>
<td>Proposed approach</td>
<td>97.08</td>
</tr>
</tbody>
</table>

Comparison with the state-of-the-art results on UTKinect-Action dataset.
Results: Florence3D-Action Dataset

- Total 215 action sequences: 9 actions performed (2 or 3 times) by 10 different subjects.
- Half of the subjects were used for training and the other half were used for testing.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Part Bag-of-Poses</td>
<td>82.00</td>
</tr>
<tr>
<td>Proposed approach</td>
<td>90.88</td>
</tr>
</tbody>
</table>

Recognition rates for various skeletal representations on Florence3D-Action dataset.

Comparison with the state-of-the-art results on Florence3D-Action dataset.
Thank You