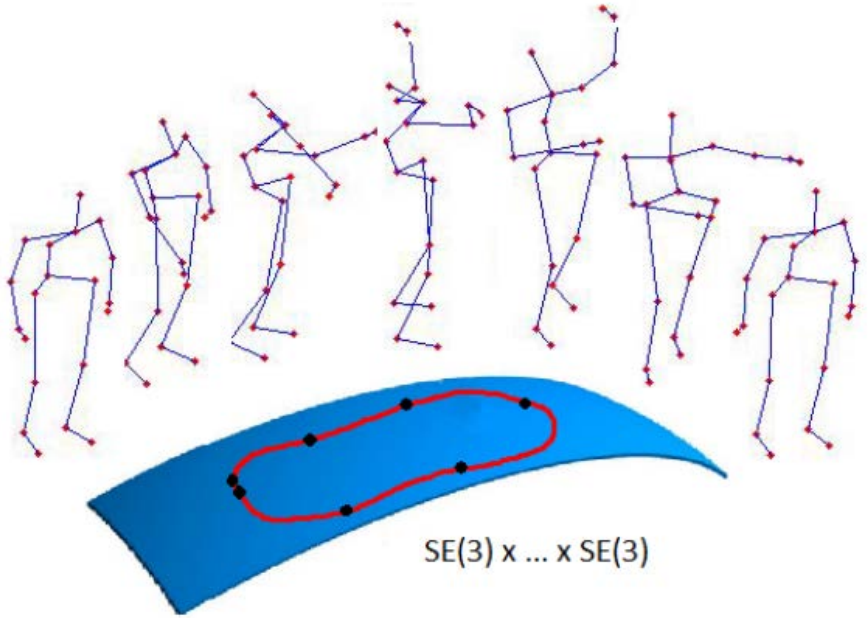
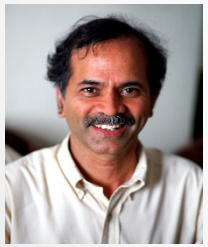
Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group





Professor Rama Chellappa

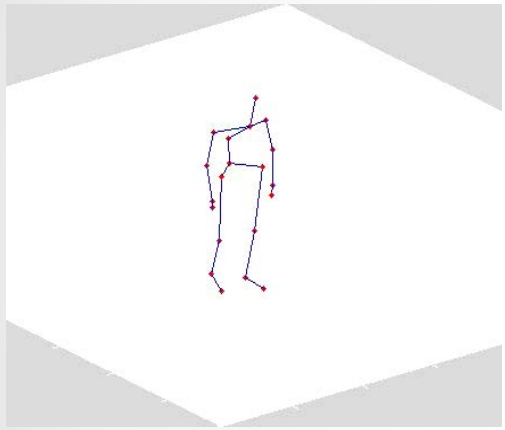
Raviteja Vemulapalli

University of Maryland, College Park.

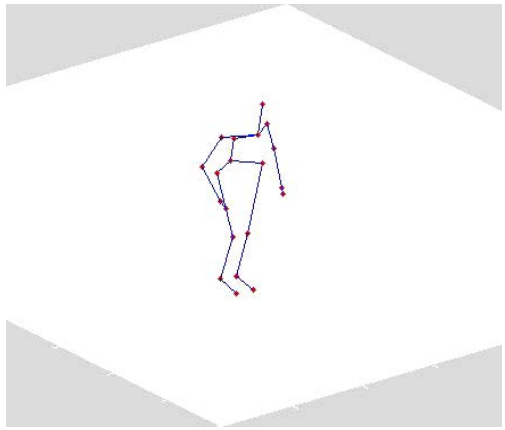
Dr. Felipe Arrate

Action Recognition from 3D Skeletal Data

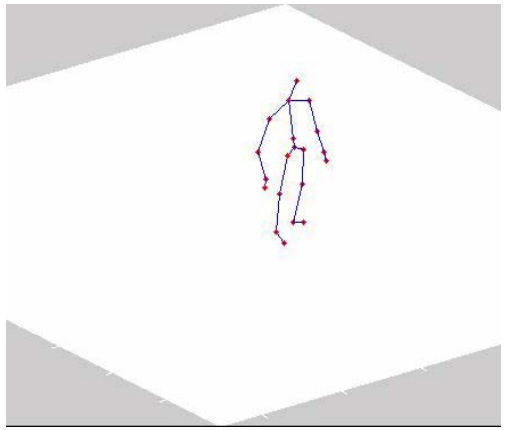
- Motivation: Humans can recognize many actions directly from skeletal sequences.



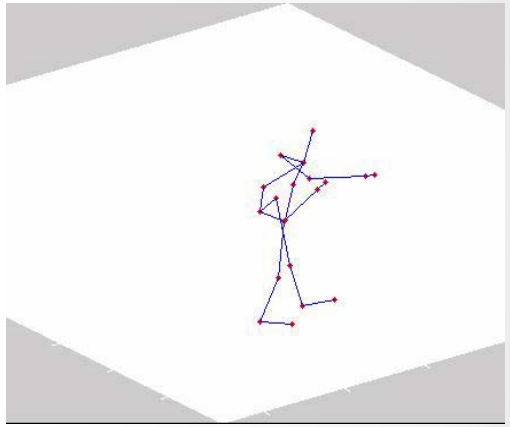
Tennis serve



Jogging



Sit down

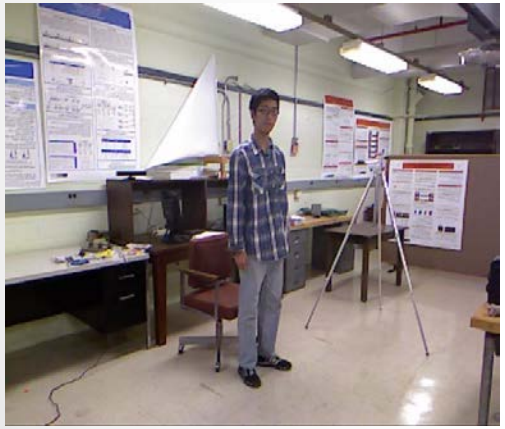


Boxing

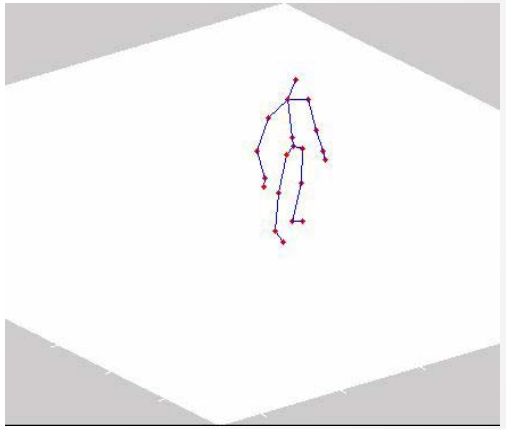
But, how do we get the 3D skeletal data?

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]



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.

L. Xia, C. C. Chen and J. K. Aggarwal, "View Invariant Human Action Recognition using Histograms of 3D Joints ", In CVPRW, 2012.

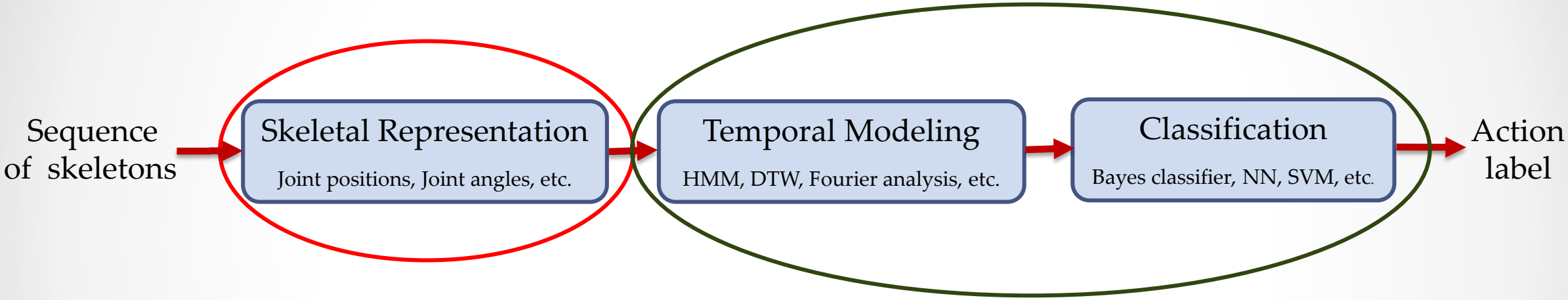
Applications

Gesture-based Control

Elderly Care

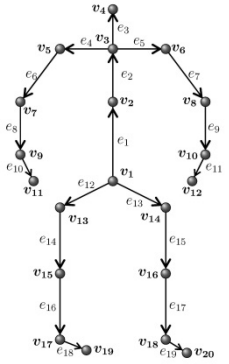
Teaching Robots

Skeleton-based Action Recognition

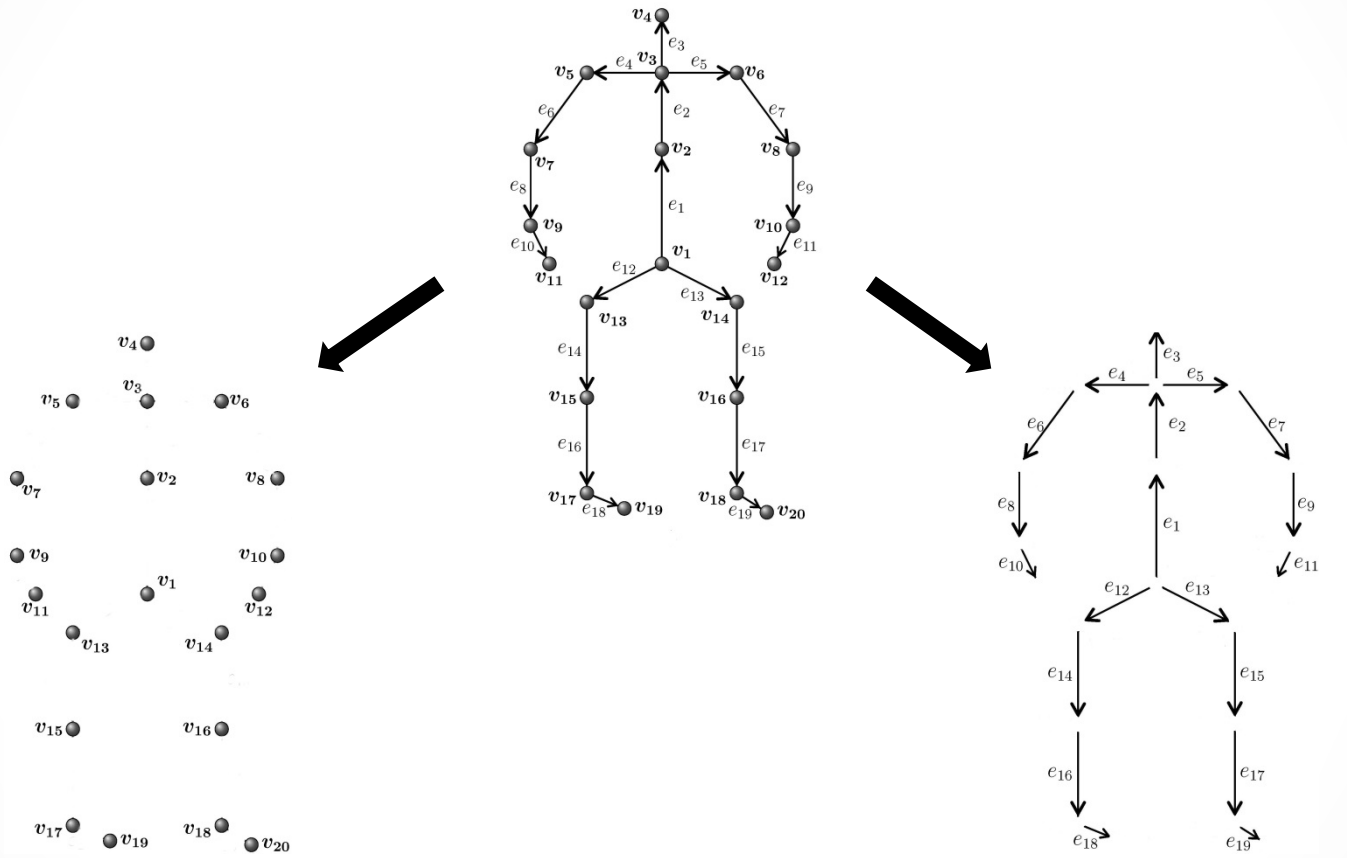


Overview of a typical skeleton-based action recognition approach.

How to represent a 3D human skeleton for action recognition ?



Human Skeleton: Points or Rigid Rods?

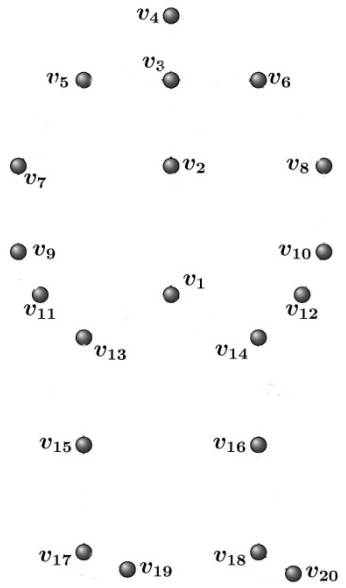


Set of points
(joints)

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.



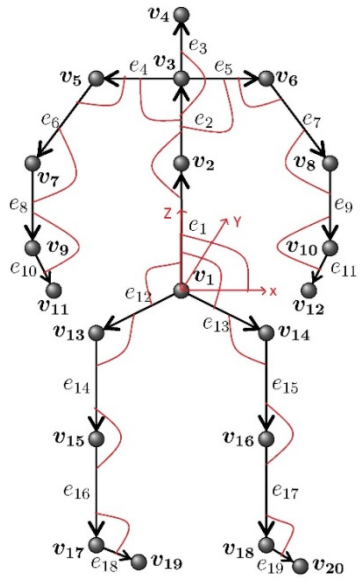
Representation:

Concatenation of the 3D coordinates of the joints.

G. Johansson, “Visual Perception of Biological Motion and a Model for its Analysis", Perception and Psychophysics, 1973.

Human Skeleton as a Set of Rigid Rods

- 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.

Proposed Representation: Motivation

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.

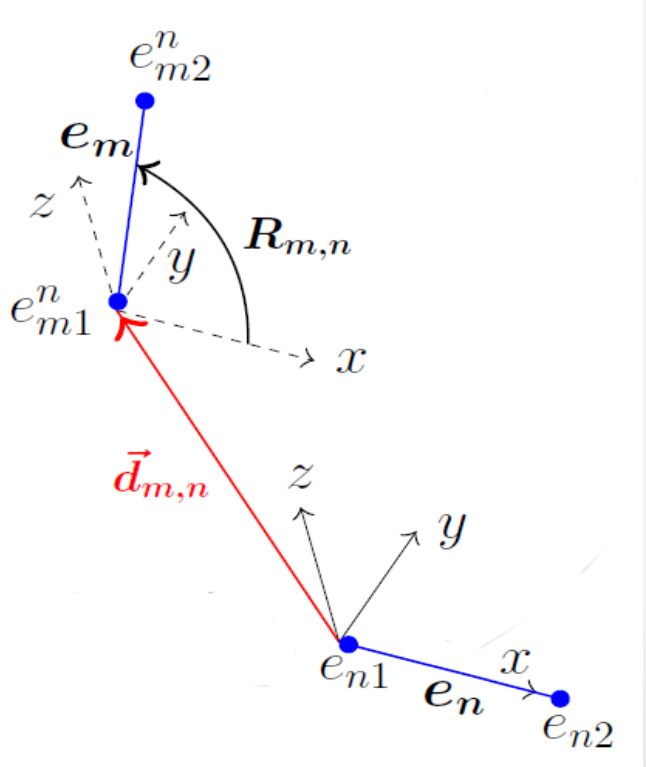
Relative 3D Geometry between Body Parts

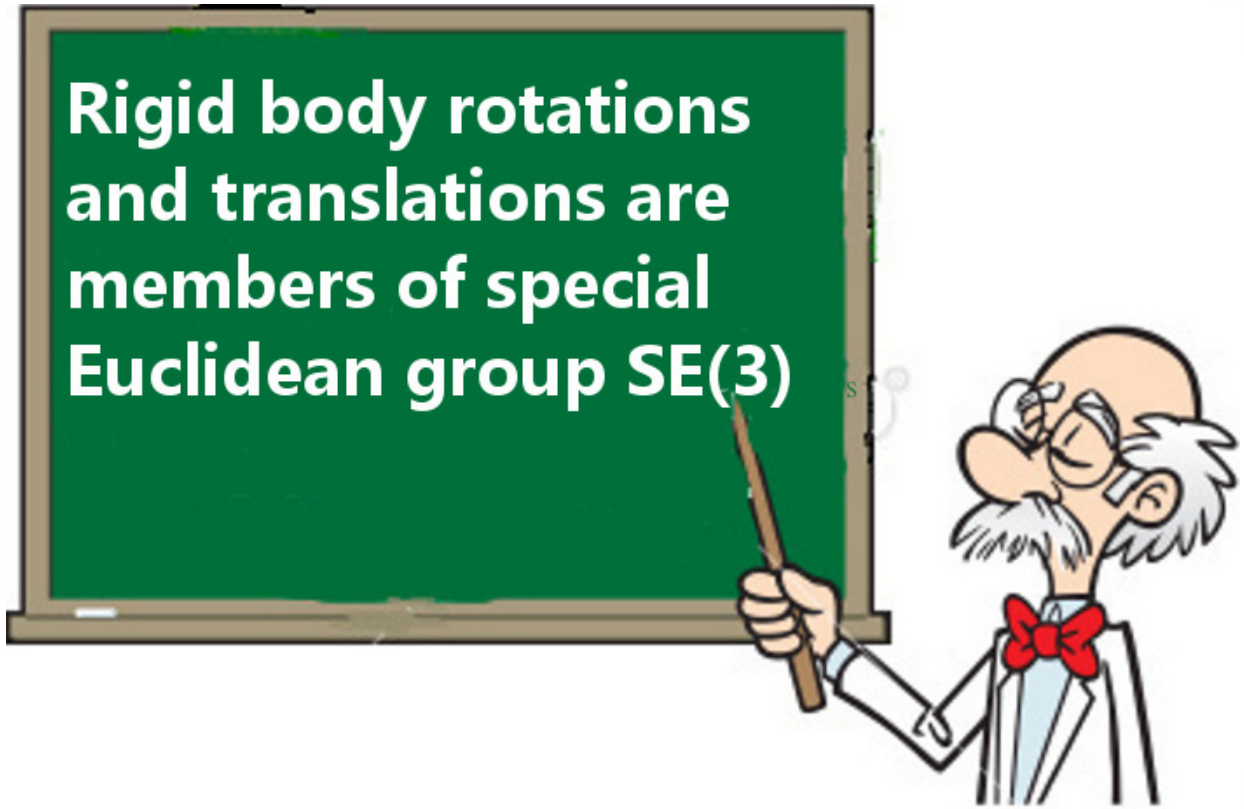
- We describe the relative geometry between two rigid body parts (e_m, e_n) at time instance t using the rotation $R_{m,n}(t)$ and the translation $\vec{d}_{m,n}(t)$ (measured in the local coordinate system attached to e_n) required to take e_n to the position and orientation of e_m.

$$\begin{bmatrix} e^n_{m1}(t) & e^n_{m2}(t) \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} R_{m,n}(t) & \vec{d}_{m,n}(t) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} e_{n1}(t) & s_{m,n} * e_{n2}(t) \\ 1 & 1 \end{bmatrix}$$

Rotation and translation vary with time.

Scaling factor: Independent of time since lengths of the body parts do not change with time.





R. M. Murray, Z. Li, and S. S. Sastry, "A Mathematical Introduction to Robotic Manipulation", CRC Press, 1994.

Special Euclidean Group *SE(3)*

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

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

  where $\vec{d} \in R^3$ and R is a 3×3 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 $exp_{SE(3)}: \mathfrak{se}(3) \rightarrow SE(3)$ and the logarithm map $log_{SE(3)}: SE(3) \rightarrow \mathfrak{se}(3)$ are given by

$$exp_{SE(3)}(B) = \boldsymbol{e}^B,$$

$$log_{SE(3)}(P) = \boldsymbol{log}(P),$$

  where $\boldsymbol{e}$ and $\boldsymbol{log}$ denote the usual matrix exponential and logarithm.

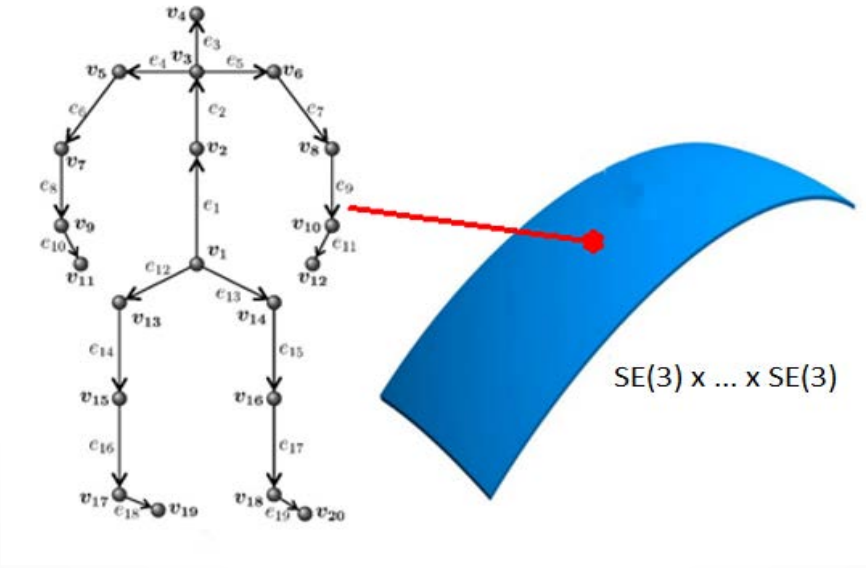
Proposed Skeletal Representation

- Human skeleton is described using the relative 3D geometry between all pairs of body parts.

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

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 ×.

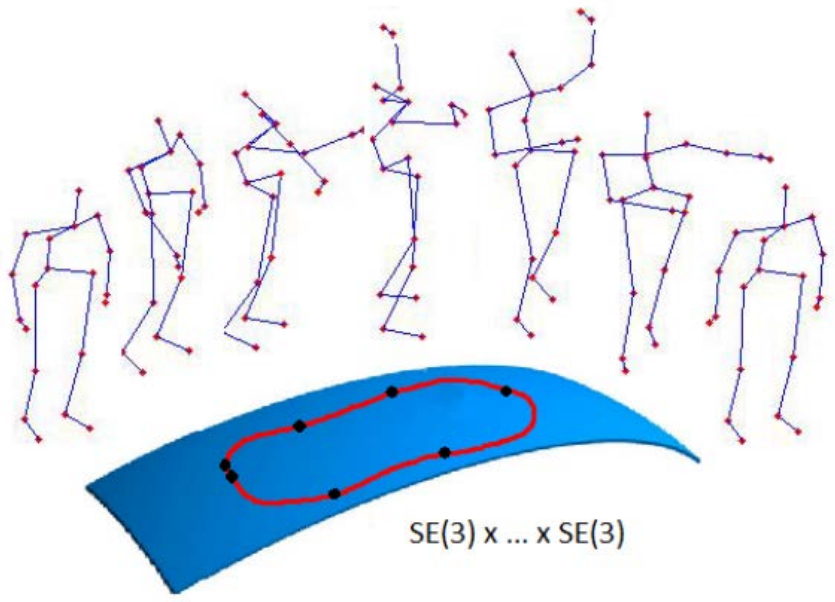


Proposed Action Representation

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

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

Point in *SE(3)* × ⋯ × *SE(3)* representing the skeleton at time instance *t*.



Proposed Action Representation

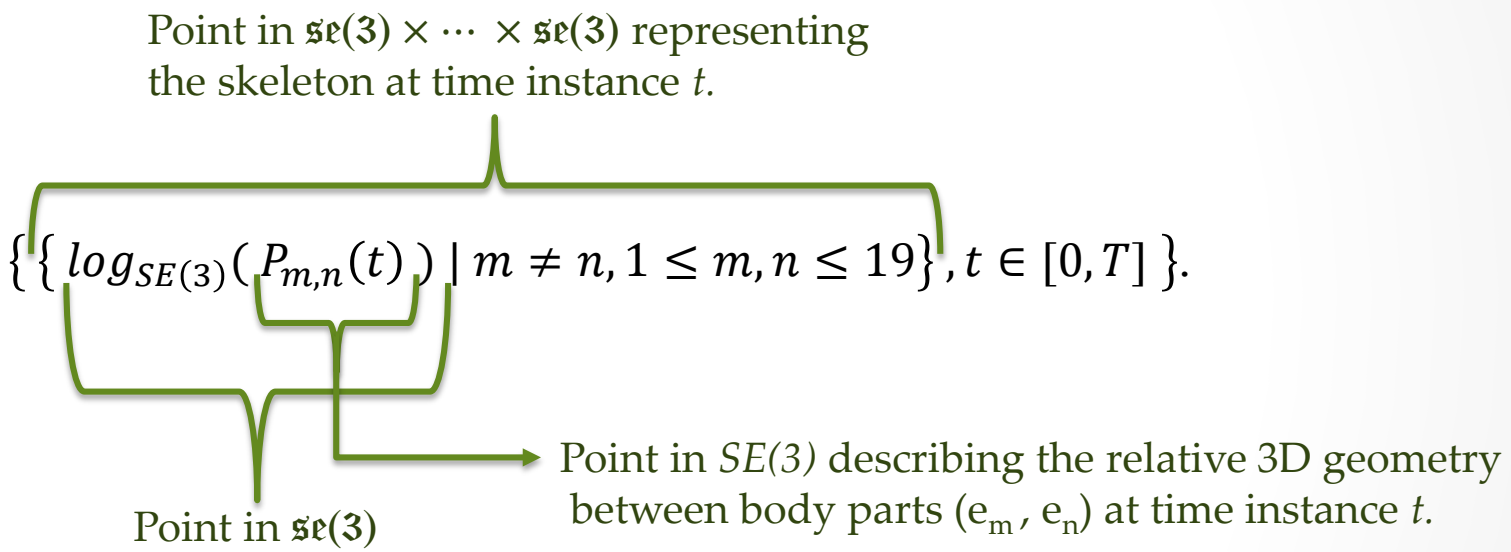
- Classification of the curves in $SE(3) \times \cdots \times SE(3)$ into different action categories is a difficult task due to the non-Euclidean nature of the space.

- 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 $\mathfrak{se}(3) \times \cdots \times \mathfrak{se}(3)$.



- 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.

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.

J. Wang, Z. Liu, Y. Wu and J. Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras", In CVPR, 2012.

Computation of Nominal Curves using DTW

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

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 $t = 0, 1, \ldots, T$.

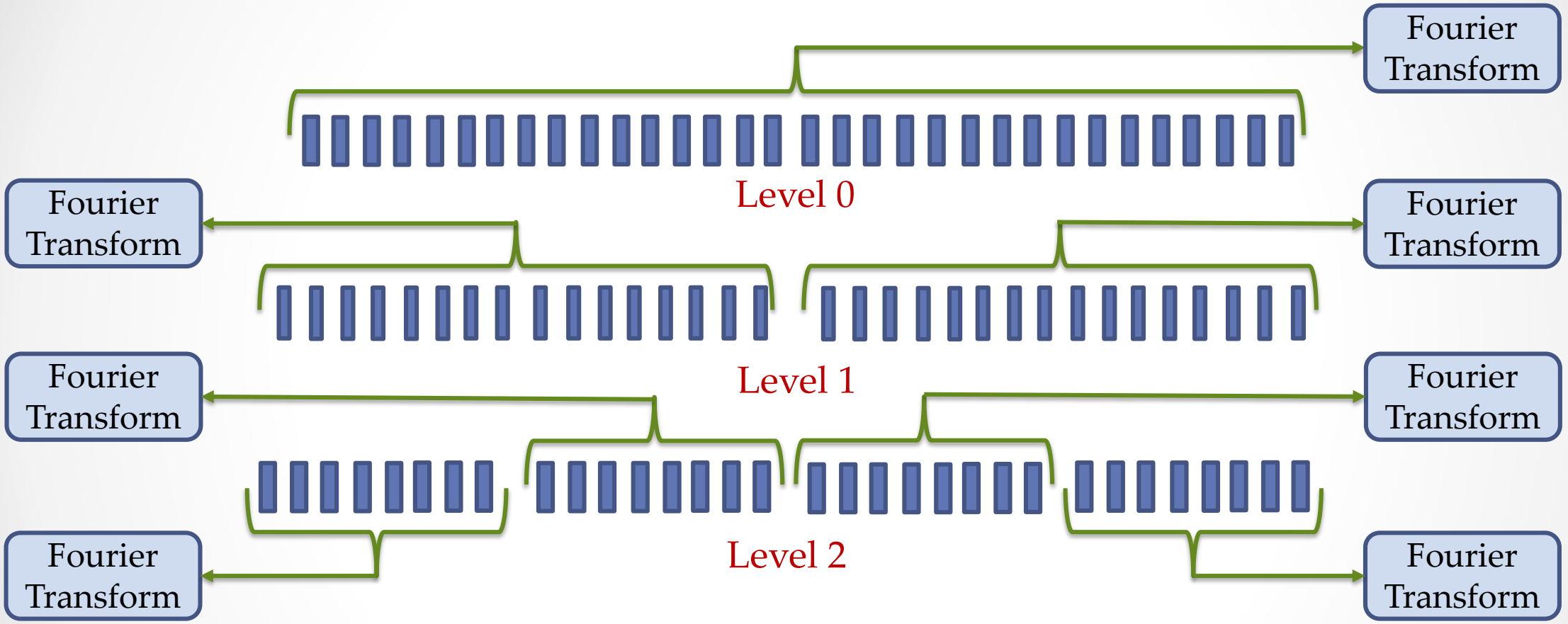
Initialization: $\mathfrak{C}(t) = \mathfrak{C}_1(t)$, iter = 0.
while iter $< max$
  Warp each curve $\mathfrak{C}_j(t)$ to the nominal curve $\mathfrak{C}(t)$ using DTW with squared Euclidean distance to get a warped curve $\mathfrak{C}_j^w(t)$.

  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

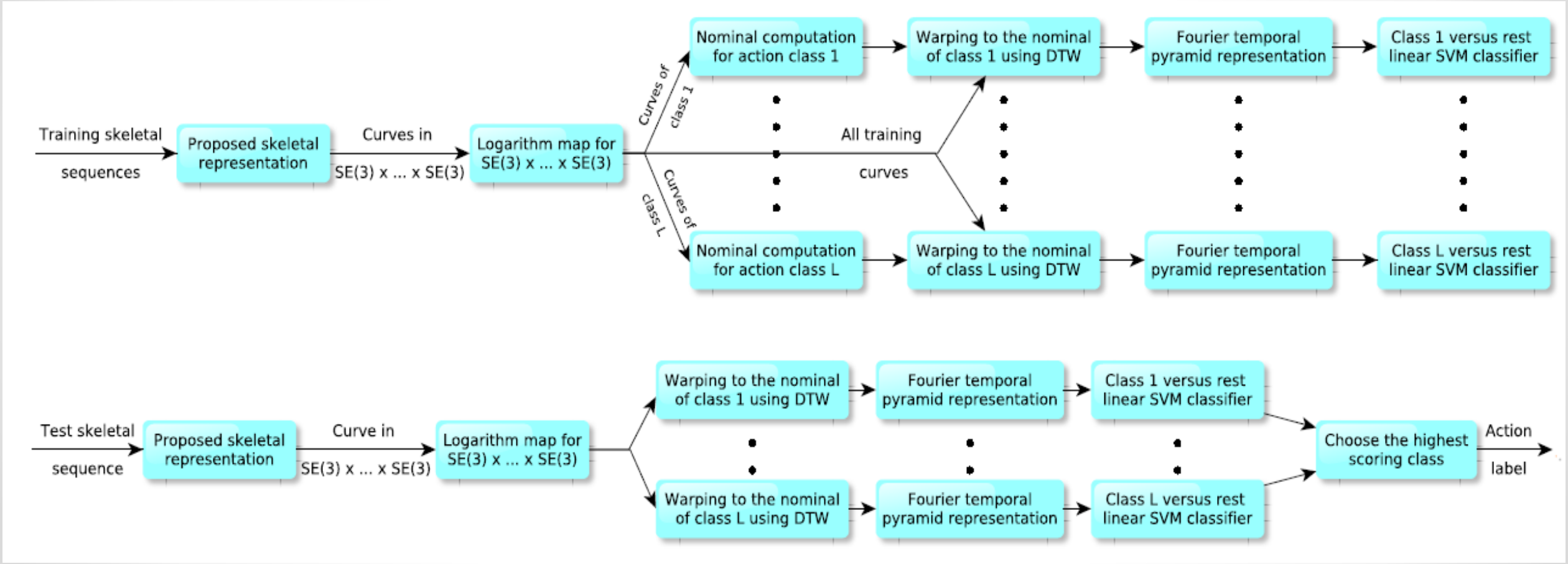
Fourier Temporal Pyramid Representation



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

J. Wang, Z. Liu, Y. Wu and J. Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras", In CVPR, 2012.

Overview of the Proposed Approach



Experiments: Datasets

MSR-Action3D dataset

- Total 557 action sequences
- 20 actions
- 10 subjects

W. Li, Z. Zhang, and Z. Liu, "Action Recognition Based on a Bag of 3D Points", In CVPR Workshops, 2010.

UTKinect-Action dataset

- Total 199 action sequences
- 10 actions
- 10 subjects

L. Xia, C. C. Chen, and J. K. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints", In CVPR Workshops, 2012.

Florence3D-Action dataset

- Total 215 action sequences
- 9 actions
- 10 subjects

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 CVPR Workshops, 2013.

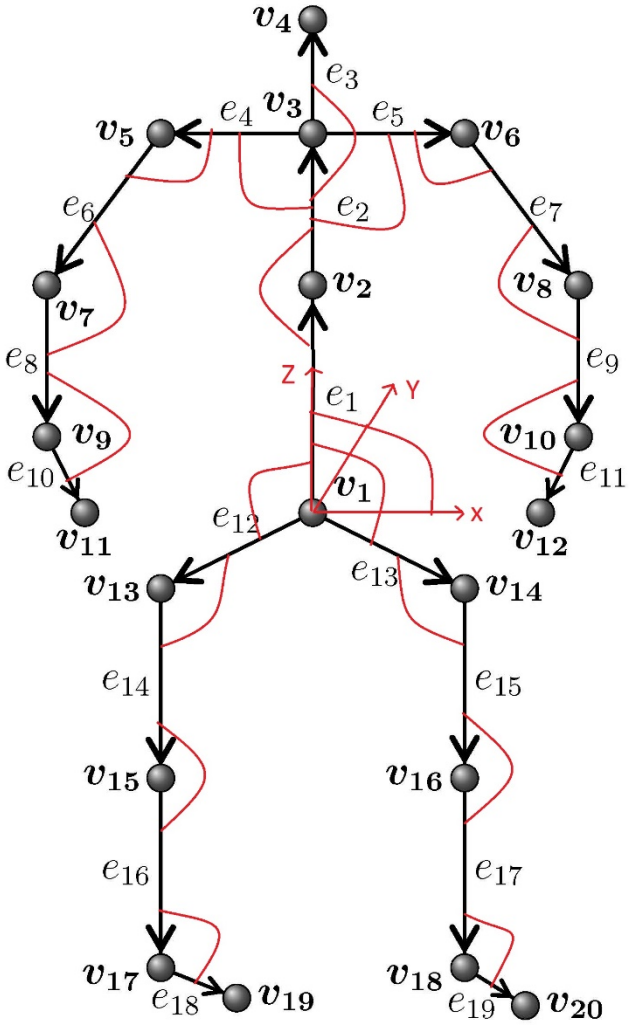
Alternative Representations for Comparison

Joint positions (JP):

Concatenation of the 3D coordinates of the joints.

Pairwise relative positions of the joints (RJP):

Concatenation of the 3D vectors $\overrightarrow{v_i v_j},\ 1 \leq i < j \leq 20$.



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.

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side boxing	Pickup & throw

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.

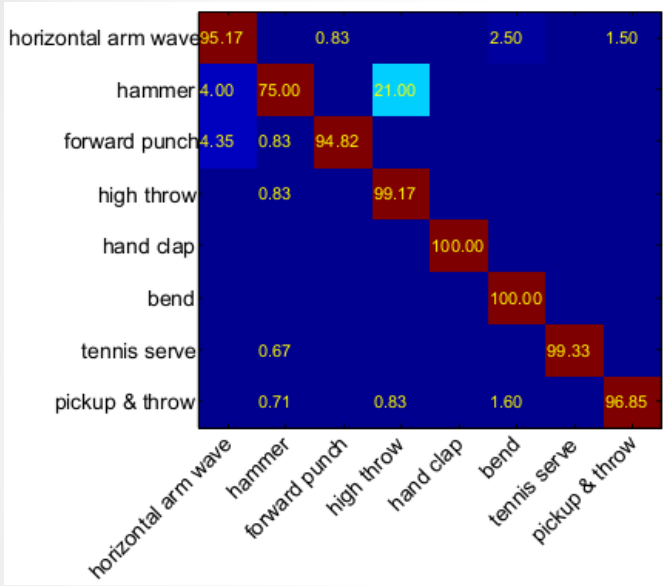
Dataset	JP	RJP	JA	BPL	Proposed
AS1	91.65	92.15	85.80	83.87	**95.29**
AS2	75.36	79.24	65.47	75.23	**83.87**
AS3	94.64	93.31	94.22	91.54	**98.22**
Average	87.22	88.23	81.83	83.54	**92.46**

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

Approach	Accuracy
Eigen Joints	82.30
Joint angle similarities	83.53
Spatial and temporal part-sets	90.22
Covariance descriptors on 3D joint locations	90.53
Random forests	90.90
Proposed approach	**92.46**

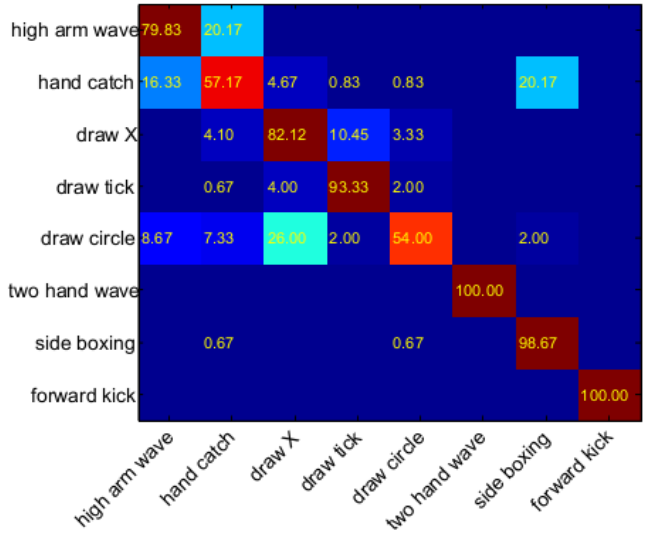
Comparison with the state-of-the-art results on MSR-Action3D dataset.

MSR-Action3D Confusion Matrices



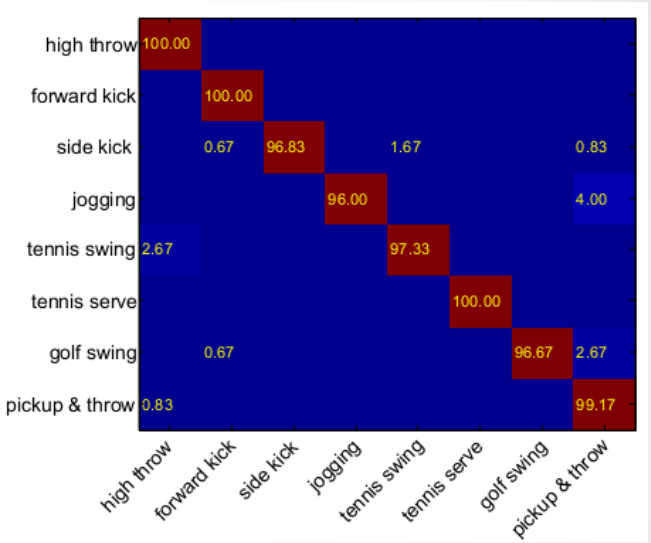
Action set 1 (AS1)

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

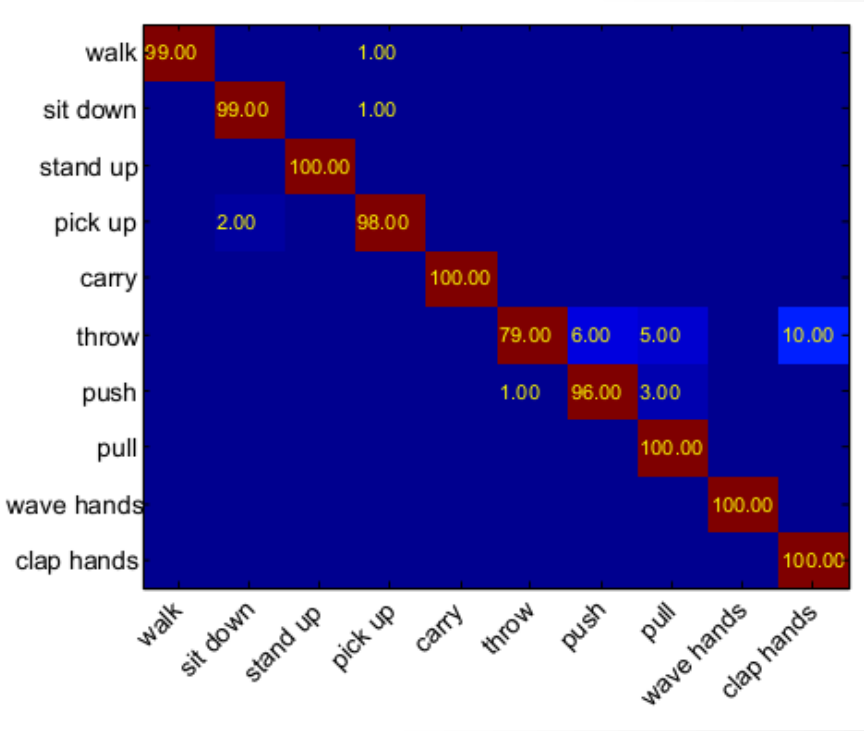
- Total 199 action sequences: 10 actions performed (2 times) by 10 different subjects.
- Half of the subjects were used for training and the other half were used for testing.

JP	RJP	JA	BPL	Proposed
94.68	95.58	94.07	94.57	**97.08**

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

Approach	Accuracy
Random forests	87.90
Histograms of 3D joints	90.92
Proposed approach	**97.08**

Comparison with the state-of-the-art results on UTKinect-Action dataset.



Results: Florence3D-Action Dataset

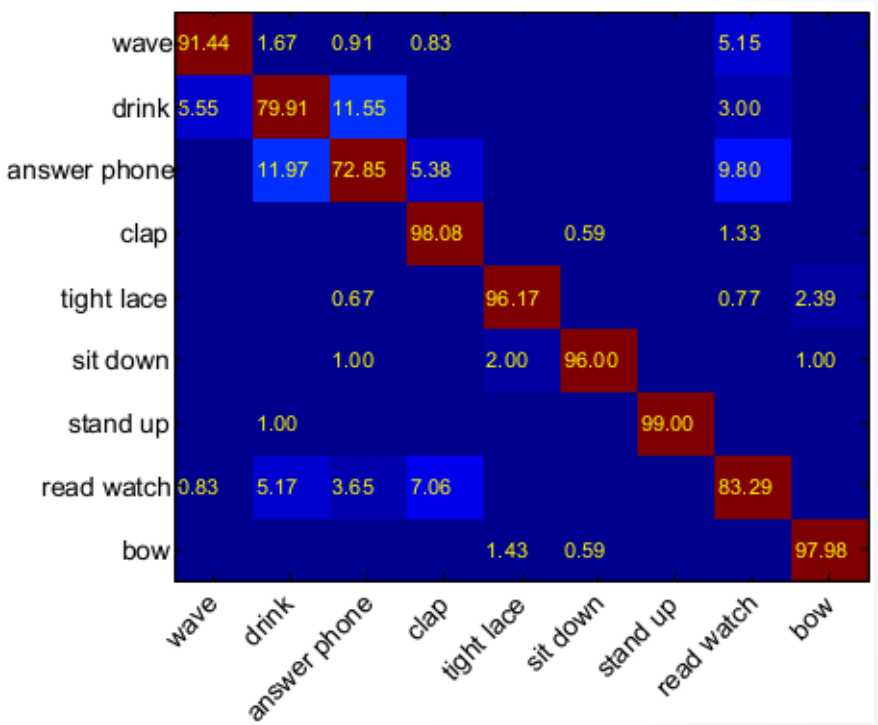
- Total 215 action sequences: 9 actions performed (2 or 3 times) by 10 different subjects.
- Half of the subjects were used for training and the other half were used for testing.

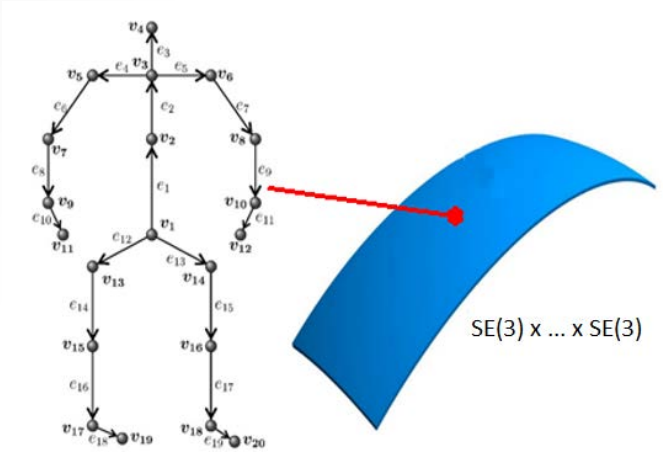
JP	RJP	JA	BPL	Proposed
85.26	85.2	81.36	80.80	**90.88**

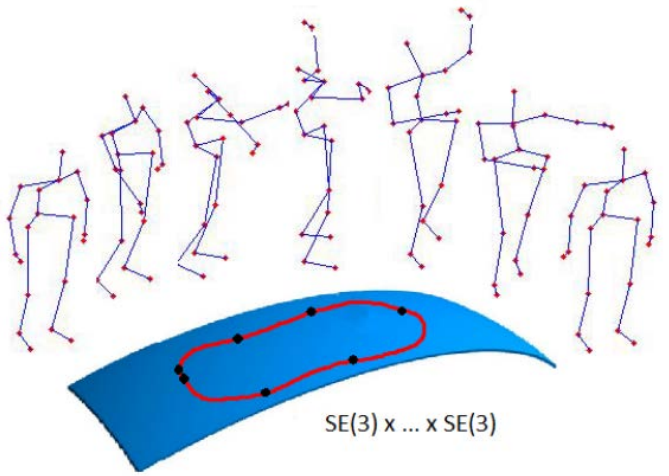
Recognition rates for various skeletal representations on Florence3D-Action dataset.

Approach	Accuracy
Multi-Part Bag-of-Poses	82.00
Proposed approach	**90.88**

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







Thank You