Computer Vision and Image Understanding 152 (2016) 155–166

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

R3DG features: Relative 3D geometry-based skeletal representations for human action recognition

Raviteja Vemulapalli[a,*], Felipe Arrate[b], Rama Chellappa[a]

[a] Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742, United States
[b] Advanced Center for Electrical and Electronic Engineering, Universidad Técnica Federico Santa Maria, Valparaiso, Chile

ARTICLE INFO

Article history:
Received 26 October 2015
Revised 6 February 2016
Accepted 12 April 2016
Available online 13 April 2016

Keywords:
Action recognition
Skeletal representations
Lie groups
Special orthogonal group
Special Euclidean group
Quaternions
Dual quaternions

ABSTRACT

Recently introduced cost-effective depth sensors coupled with real-time skeleton extraction algorithms 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 the human skeleton. In this paper, we introduce and evaluate a new family of skeletal representations for human action recognition, which we refer to as R3DG features. The proposed representations explicitly model the 3D geometric relationships between various body parts using rigid body transformations, i.e., rotations and translations in 3D space. Using the proposed skeletal representations, human actions are modeled as curves in R3DG feature spaces. Finally, we perform action recognition by classifying these curves using a combination of dynamic time warping, Fourier temporal pyramid representation and support vector machines. Experimental results on five benchmark action datasets show that the proposed representations perform 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, human computer interaction, robotics, health care, etc. In the past few decades, several approaches have been proposed for recognizing human actions from monocular RGB video sequences (Aggarwal and Ryoo, 2011; Turaga et al., 2008). 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 a 3D space. Hence, despite significant research efforts over the past few decades, human action recognition still remains a challenging problem.

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 (Zatsiorsky, 1997). So, if we can reliably extract the human skeleton, action recognition can be performed by classifying its temporal evolution. Using skeletal data for action recognition has several advantages such as ease of interpretability, low processing time, fast/cheap transmission and storage, etc. Skeletal data makes it easier to analyze which part of the body is playing a major role in discriminating one action against the other, and allows us to correlate this with human interpretation of motion. Interpretability is an important factor in various applications such as exercise monitoring, human computer interaction, post-surgery rehabilitation, etc. Skeletons provide a compact low-dimensional representation that can be stored easily, transmitted and processed quickly. Storage and transmission are critical in applications where the recognition module runs on a central server.

Unfortunately, extracting the human skeleton from monocular RGB videos is a very difficult task (Moeslund et al., 2006). Sophisticated motion capture systems can be used to get 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 availability 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 infer human skeletons. Recently, a quick method was proposed in Shotton et al. (2011) to accurately estimate the 3D positions of skeletal joints using a single depth image. These recent advances have generated 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

* Corresponding author. Tel: +1 240-338-8702.
E-mail address: raviteja@umd.edu (R. Vemulapalli).

http://dx.doi.org/10.1016/j.cviu.2016.04.005
1077-3142/

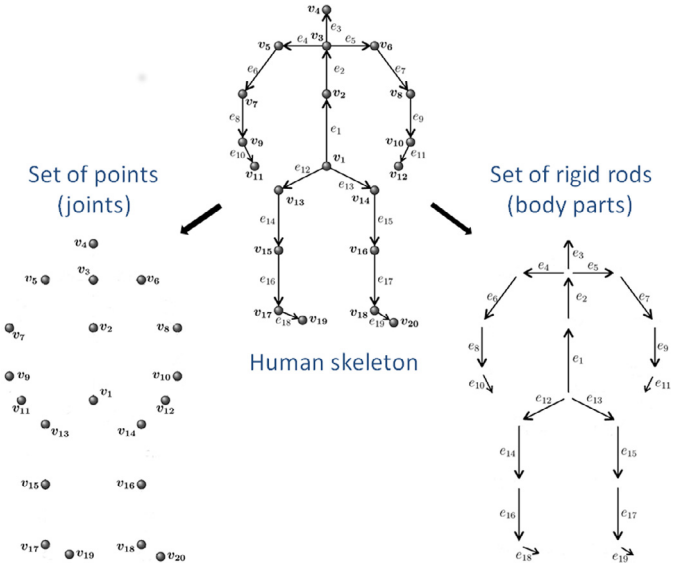

Fig. 1. Two views of the human skeleton: left - set of points, right - set of rigid rods.

and body part-based approaches. Inspired by the moving lights display experiment of Johansson (1973), joint-based approaches consider the human skeleton as a set of points (Fig. 1, left). These approaches try to model the motion of either individual joints or combinations of multiple joints using various features like joint positions (Hussein et al., 2013; Lv and Nevatia, 2006; Reyes et al., 2011; Sheikh et al., 2005), joint orientations with respect to a fixed root node (Xia et al., 2012; Shao and Li, 2013), pairwise relative joint positions (Wang et al., 2012b; Wei et al., 2013; Yang and Tian, 2014a), etc. On the other hand, motivated by the 3D-shape representations of Marr and Nishihara (1978), body part-based approaches consider the human skeleton as a connected set of rigid segments (Fig. 1, right). These approaches either model the temporal evolution of individual body parts (Yacoob and Black, 1998) or focus on directly-connected pairs of body parts and model the temporal evolution of joint angles (Gavrila and Davis, 1995, Ofli et al., 2014; Ohn-bar and Trivedi, 2013).

In this paper, we introduce a new family of body part-based skeletal representations for recognizing human actions. 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 (for example, 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 representations.

Given two rigid body parts, their relative geometry can be described using the rigid body transformation (rotation and translation) required to take one body part to the position and orientation of the other. Hence, we use the rigid body transformations between all pairs of body parts to represent the human skeleton. Rigid body transformations in 3D space can be mathematically represented in various different ways using the special orthogonal group $SO(3)$, quaternions, the special Euclidean group $SE(3)$, and dual quaternions. Using these mathematical representations, we introduce a family of relative 3D geometry-based skeletal representations for action recognition, which we refer to as R3DG features.

One of the major issues while working with skeletal-data is scale variation. This can be handled by normalizing all the skeletons (without changing the joint angles) such that their body part lengths are equal to the corresponding lengths of a fixed reference skeleton. Note that a full 3D rigid body transformation includes both rotation and translation. Interestingly, while the relative translations between various body parts vary with this scale normalization, the relative rotations do not change (even if the body parts are not directly connected by a joint). Hence, we can get scale-invariant skeletal representations by using only the relative rotations between different body parts. In this work, we experimentally show that just by using the relative 3D rotations, we can get a classification accuracy that is close to the accuracy obtained by using the full rigid body transformation-based representations computed from scale-normalized skeletons.

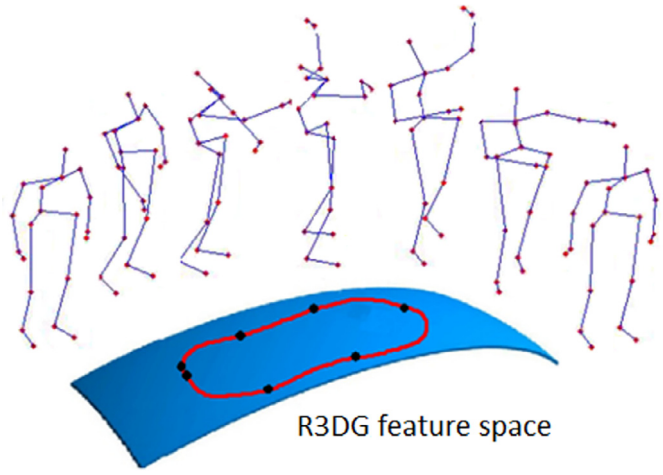

Fig. 2. Representation of an action as a curve in an R3DG feature space.

Using any of the proposed skeletal representations, human actions can be modeled as curves (Fig. 2) in an R3DG feature space, and action recognition can be performed by classifying these curves. Irrespective of the skeletal representation being used, classification of temporal sequences into different action categories is a difficult problem due to various 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 (Müller, 2007), and warp all the curves to this nominal curve. Then, we represent the warped curves using the low frequency Fourier temporal pyramid (FTP) representation, which was shown to be robust to noise and temporal misalignment in Wang et al. (2012b). Finally, classification is performed using a support vector machines (SVM) classifier with the FTP representation.

A preliminary version of this work appeared in Vemulapalli et al. (2014). The following are the additional contributions of this work compared to the earlier version:

1. We introduce and evaluate a new family of body part-based 3D skeletal representations for human action recognition. The proposed representations explicitly model the relative geometry between various body parts using 3D rigid body transformations. While Vemulapalli et al. (2014) used $SE(3)$ to represent 3D rigid body transformations, we consider three other alternative representations based on $SO(3)$, quaternions and dual quaternions. To the best of our knowledge, $SE(3)$ and dual quaternions have not been explored before in the context of skeleton-based human action recognition.
2. We introduce scale-invariant R3DG features that use only the relative 3D rotations between various body parts.
3. We experimentally show that the proposed representations outperform many existing skeletal representations by evaluating them on five action datasets (only three were used in Vemulapalli et al. (2014)). We also include comparisons with the skeletal quad descriptor (Evangelidis et al., 2014) and the

relational pose features (Yao et al., 2012; Yun et al., 2012), which were missing in Vemulapalli et al. (2014).
4. To make the paper self-contained, we include Sections 3 and 4 which provide relevant background information on $SO(3)$, $SE(3)$, quaternions, and dual quaternions.

Organization: We provide a brief overview of existing skeleton-based action recognition approaches in Section 2, and briefly discuss $SO(3)$, $SE(3)$, quaternions, and dual quaternions in Sections 3 and 4. We introduce the proposed family of R3DG features in Section 5, and describe the temporal modeling and classification approach in Section 6. Experimental results and conclusions are presented in Sections 7 and 8, respectively.

Notations: We use $\mathcal{R}^n$ to represent the n-dimensional real vector space, and I to denote an identity matrix of appropriate size. Vectors are represented with an arrow on top. We use $\vec{0}$ to represent the zero vector in $\mathcal{R}^3$. The determinant of a matrix R is represented using $\det(R)$ and the ℓ_2-norm of a vector $\vec{v}$ is denoted by $\|\vec{v}\|_2$. We use $\otimes$ to represent the direct product between groups and $\oplus$ to represent the direct sum between vector spaces.

2. Relevant work

In this section, we provide a brief overview of various existing skeleton-based human action recognition approaches. Various depth map-based action recognition approaches (Chen et al., 2015; 2014; Hu et al., 2015; Kong and Fu, 2015; Lu et al., 2014; Oreifej and Liu, 2013; Rahmani et al., 2014; Wang et al., 2012a; Xia and Aggarwal, 2013; Yang and Tian, 2014b; Yang et al., 2012; Zhang and Tian, 2015; Zhu et al., 2014) have also been proposed in the recent past, which use various features extracted from the 3D depth data. Since the focus of this work is on skeleton-based action recognition, we refer the readers to Chen et al. (2013); Ye et al. (2013) for a review of depth map-based recognition approaches.

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

2.1. Joint-based approaches

A set of 13 joint trajectories in XYZT space was used to represent human actions in Sheikh et al. (2005), and their affine projections were compared using a subspace angles-based similarity measure. In Lv and Nevatia (2006), the trajectories of individual joints and groups of joints were modeled using hidden Markov models (HMMs). Each HMM was considered as a weak classifier, which were then combined using AdaBoost. HMMs were also used in Gu et al. (2010) to model the joint trajectories of whole body, upper body and lower body separately for performing action recognition.

The 3D joint locations were combined with silhouette-based features in Chaaraoui et al. (2013), and their temporal evolutions were compared using dynamic time warping (DTW). Dynamic time warping was also used in Reyes et al. (2011) for comparing the sequences of joint positions. Instead of giving equal weight to all the joints in the DTW distance computation, a feature weighting approach was used in Reyes et al. (2011), where each joint was assigned its own weight. In Hussein et al. (2013), the temporal evolutions of joint locations were modeled using a temporal hierarchy of covariance features, and action recognition was performed using an SVM. In Gowayyed et al. (2013), the 3D trajectory of each joint was projected onto three Cartesian planes to get three 2D trajectories. Each 2D trajectory was represented using the histogram of displacements between consecutive points. The histograms from all the joints were concatenated to get the final feature representation, which was classified using an SVM. Recently, recurrent neural networks were used in Du et al. (2015) for modeling the temporal dynamics of skeletal joints.

A view invariant representation of human skeleton was obtained in Xia et al. (2012) by quantizing the 3D joint locations into histograms based on their orientations with respect to a coordinate system attached to the hip center. The temporal evolutions of this representation were modeled using HMMs. In Zanfir et al. (2013), human skeletons were represented using 3D joint positions, their first and second order derivatives, i.e., joint velocities and accelerations, and a nearest neighbor-based approach was used to perform low-latency action recognition. In Shao and Li (2013), one of the joints was selected as a root joint, and all the remaining joints were represented using their orientations with respect to a coordinate system attached to the root joint. The temporal evolutions of this representation were compared using dynamic time warping.

In Wang et al. (2012b); Wei et al. (2013), pairwise relative positions of the joints were used to represent the human skeleton, and the temporal evolutions of this representation were modeled using wavelets (Wei et al., 2013) and low-frequency Fourier coefficients (Wang et al., 2012b). A similar skeletal representation was also used in Wang and Wu (2013), where a discriminative learning-based temporal alignment method was used for comparing temporal sequences.

In Ellis et al. (2013), the human skeleton was represented using distances between all pairs of joints in the current frame, distances between all pairs of joints in the current frame and the previous frame, distances between all pairs of joints in the current frame and the first frame of the sequence. Action recognition was then performed using a logistic regression-based approach. In Yang and Tian (2014a), the human skeleton was represented using relative joint positions, temporal displacements of the joints, and offsets of the joints with respect to the initial frame. Action classification was then performed using the Naive-Bayes nearest neighbor rule in a low-dimensional space obtained using principal component analysis (PCA). A similar representation was also used in Zhu et al. (2013) along with random forests.

A local skeleton descriptor, referred to as *skeletal quad*, was introduced in Evangelidis et al. (2014), which encodes the relative position of joint quadruples. This descriptor represents a set of four joints using the coordinates of third and fourth joints in a coordinate system with the first joint as the origin and the second joint as (1, 1, 1). These skeletal quads were combined with Fisher vectors (Jaakkola and Haussler, 1998) and a linear SVM to perform action recognition. An interesting aspect of this descriptor is that it can be used to represent the relative 3D geometry between two body parts (since two body parts can be considered as four joints). However, the main difference between the skeletal quad descriptor and the proposed R3DG features is that while the proposed features directly use the translation and rotation between body parts, the skeletal quad descriptor encodes this information indirectly using the joint coordinates.

In Wang et al. (2013), human skeleton was divided into five parts and each part was represented using the coordinates of the joints that belonged to the part. Then, a dictionary of pose templates was learned for each body part, and these templates were used to obtain a quantized representation of part poses. The authors further defined spatial-part-sets to capture the spatial configurations of multiple body parts, and temporal-part-sets to capture the joint pose evolutions of multiple body parts. Finally,

the bag-of-words model was used to get the action representation, which was classified using one-vs-one intersection kernel SVM.

Most of the above mentioned approaches use either the joint positions or the relative joint positions to represent the human skeleton. Different from these, Müller et al. (2005) introduced various types of relational pose features that describe geometric relations between specified joints of the skeleton, and used them successfully for indexing and retrieval of motion capture data. Similar features were later used in Yao et al. (2011); 2012); Yun et al. (2012) for skeleton-based human action recognition.

2.2. Part-based approaches

In Yacoob and Black (1998), the human body was divided into five different parts, and human actions were represented using the motion parameters of individual 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 basis actions, and classification was performed by comparing the PCA coefficients. In Chaudhry et al. (2013), human skeletons were hierarchically divided into smaller parts and each body part was represented using certain bio-inspired shape features. The temporal evolutions of these bio-inspired features were modeled using linear dynamical systems.

In Gavrila and Davis (1995), human skeletons were represented using 3D joint angles, and the temporal evolutions of this representation were compared using DTW. While Campbell and Bobick (1995) represented human actions as curves in low-dimensional phase spaces related to joint angles, Ohn-bar and Trivedi (2013) represented human actions using pairwise affinities between joint angle trajectories. In Sung et al. (2012), human skeletons were represented using joint angle quaternions. These skeletal features were augmented with RGB and depth-based HOG features, and a maximum entropy Markov model was used for action detection. In Ofli et al. (2014), a set of few informative skeletal joints was selected at each time instance based on highly interpretable measures such as mean and variance of joint angles, angular velocity of the joints, etc. Human actions were represented as sequences of these informative joints, which were compared using the normalized edit distance.

3. Lie groups

In this section, we discuss two matrix Lie groups that are of interest to us, namely the special orthogonal group $SO(3)$ and the special Euclidean group $SE(3)$. While $SO(3)$ is commonly used for representing 3D rotations, $SE(3)$ is used for representing 3D rigid body transformations. We refer the readers to Hall (2003) for a detailed introduction to general Lie groups, and Murray et al. (1994) for further details on $SO(3)$ and $SE(3)$.

A Lie group G is a group that is also a smooth manifold. The tangent space $\mathfrak{g}$ at the identity element e of G is referred to as the Lie algebra of G. A matrix Lie group is a Lie group of $n \times n$ invertible matrices with the usual matrix multiplication and inversion as the group multiplication and inversion operations, and the $n \times n$ identity matrix as the group identity element.

The mapping from a Lie algebra to the corresponding Lie group, referred to as the Lie exponential map, is given by $\exp_G(\vec{u}) = \gamma_{\vec{u}}(1)$, where $\gamma_{\vec{u}} : \mathcal{R} \rightarrow G$ is the unique one-parameter subgroup of G whose tangent vector at the identity element e is equal to $\vec{u} \in \mathfrak{g}$. The inverse of exponential map is known as logarithm map, and is denoted by $\log_G$. Fig. 3 gives an illustration of the Lie exponential and logarithms maps. In the case of matrix Lie groups, the Lie exponential and logarithm maps are given by

$$\exp_G(\vec{u}) = \mathbf{e}^{\vec{u}}, \quad \log_G(g) = \mathbf{log}(g), \tag{1}$$

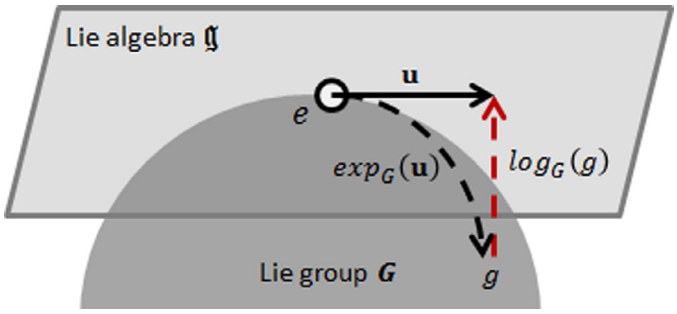

Fig. 3. Illustration of the exponential and logarithm maps between a Lie group G and its Lie algebra $\mathfrak{g}$.

where $\mathbf{e}$ and $\mathbf{log}$ represent the usual matrix exponential and logarithm respectively.

3.1. Special orthogonal group SO(3)

The special orthogonal group $SO(3)$ is a three dimensional matrix Lie group formed by the set of all 3×3 matrices R that satisfy the following constraints:

$$R^\top R = I, \ \det(R) = 1. \tag{2}$$

The Lie algebra of $SO(3)$, denoted by $\mathfrak{so}(3)$, is the three dimensional vector space spanned by the set of all 3×3 skew symmetric matrices. For any element

$$A = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix} \in \mathfrak{so}(3), \tag{3}$$

its vector representation is given by $\text{vec}(A) = [a_1, a_2, a_3]$. Since $SO(3)$ is a matrix Lie group, the Lie exponential and logarithm maps between $SO(3)$ and $\mathfrak{so}(3)$ are given by

$$\exp_{SO(3)}(A) = \mathbf{e}^A, \quad \log_{SO(3)}(R) = \mathbf{log}(R). \tag{4}$$

The logarithm map is not unique in the case of $SO(3)$. In this work, we use the $\mathbf{log}(R)$ with the smallest norm.

Elements of $SO(3)$ are commonly used to represent 3D rotations. Let $\vec{z}'$ be a 3D point obtained by rotating $\vec{z} \in \mathcal{R}^3$ by an angle θ about an axis $\vec{n}$. Then, we have

$$\vec{z}' = \mathbf{e}^{skew(\theta\vec{n})}\vec{z}, \tag{5}$$

where $skew(\theta\vec{n})$ is a skew-symmetric matrix that satisfies $\text{vec}(skew(\theta\vec{n})) = \theta\vec{n}$. Hence, the matrix $\mathbf{e}^{skew(\theta\vec{n})} \in SO(3)$ represents the 3D rotation by an angle θ about an axis $\vec{n}$.

Interpolation: Various approaches have been proposed in the past for interpolation on $SO(3)$ (Park and Ravani, 1997). In this paper, we use a simple piecewise geodesic interpolation scheme. Given $R_1, \ldots, R_m \in SO(3)$ at time instances $t_1, \ldots, t_m$ respectively, we use the following curve for interpolation:

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

where $A_i = \log_{SO(3)}\left(R_i^{-1} R_{i+1}\right)$ for $i = 1, 2, \ldots, m-1$.

$SO(3) \otimes \ldots \otimes SO(3)$: We can combine multiple $SO(3)$ groups using the direct product $\otimes$ to form a new Lie group

$$SO(3)^n := SO(3) \otimes \ldots \otimes SO(3) \tag{7}$$

with the corresponding Lie algebra

$$\mathfrak{so}(3)^n := \mathfrak{so}(3) \oplus \ldots \oplus \mathfrak{so}(3). \tag{8}$$

The Lie exponential and logarithm maps for $(A_1, \ldots, A_n) \in \mathfrak{so}(3)^n$ and $(R_1, \ldots, R_n) \in SO(3)^n$ are given by

$$\exp_{SO(3)^n}(A_1,\ldots,A_n) = (\mathbf{e}^{A_1},\ldots,\mathbf{e}^{A_n}),$$
$$\log_{SO(3)^n}(R_1,\ldots,R_n) = (\mathbf{log}(R_1),\ldots,\mathbf{log}(R_n)). \quad (9)$$

Interpolation on $SO(3)^n$ can be performed by simultaneously interpolating on individual $SO(3)$.

3.2. Special Euclidean group SE(3)

The special Euclidean group $SE(3)$ is a six dimensional matrix Lie group formed by the set of all 4×4 matrices of the form

$$P(R,\vec{d}) = \begin{bmatrix} R & \vec{d} \\ 0 & 1 \end{bmatrix},\ \vec{d} \in \mathcal{R}^3,\ R \in SO(3). \quad (10)$$

The Lie algebra of $SE(3)$, denoted by $\mathfrak{se}(3)$, is the six dimensional vector space spanned by the set of all 4×4 matrices of the form

$$B = \begin{bmatrix} 0 & -a_3 & a_2 & w_1 \\ a_3 & 0 & -a_1 & w_2 \\ -a_2 & a_1 & 0 & w_3 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \quad (11)$$

The vector representation of $B \in \mathfrak{se}(3)$ is given by

$$\mathrm{vec}(B) = [a_1, a_2, a_3, w_1, w_2, w_3]. \quad (12)$$

Since $SE(3)$ is a matrix Lie group, the Lie exponential and logarithm maps between $SE(3)$ ans $\mathfrak{se}(3)$ are given by

$$\exp_{SE(3)}(B) = \mathbf{e}^B, \quad \log_{SE(3)}(P) = \mathbf{log}(P). \quad (13)$$

The logarithm map is not unique in the case of $SE(3)$. In this work, we use the $\mathbf{log}(P)$ with the smallest norm.

Elements of $SE(3)$ are commonly used to represent 3D rigid body transformations. Let $\vec{z}'$ be a 3D point obtained by transforming $\vec{z} \in \mathcal{R}^3$ using a rotation by an angle θ about an axis $\vec{n}$ followed by a translation $\vec{d}$. Then, we have

$$\begin{bmatrix} \vec{z}' \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{e}^{skew(\theta\vec{n})} & \vec{d} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \vec{z} \\ 1 \end{bmatrix}. \quad (14)$$

Hence, the matrix $\begin{bmatrix} R & \vec{d} \\ 0 & 1 \end{bmatrix} \in SE(3)$, where $R = \mathbf{e}^{skew(\theta\vec{n})}$, represents the 3D rigid body transformation composed of a rotation by an angle θ about an axis $\vec{n}$ and a translation $\vec{d}$.

Interpolation: Various approaches have been proposed in the past for interpolation on $SE(3)$ (Belta and Kumar, 2002; Zefran and Kumar, 1998). In this paper, we use a simple piecewise interpolation scheme based on screw motions (Zefran et al., 1996). Given $P_1,\ldots,P_m \in SE(3)$ at time instances $t_1,\ldots,t_m$ respectively, we use the following curve for interpolation:

$$\gamma(t) = P_i \exp_{SE(3)}\left(\frac{t-t_i}{t_{i+1}-t_i} B_i\right) \text{ for } t \in [t_i, t_{i+1}], \quad (15)$$

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

$SE(3) \otimes \ldots \otimes SE(3)$: We can combine multiple $SE(3)$ groups using the direct product $\otimes$ to form a new Lie group

$$SE(3)^n := SE(3) \otimes \ldots \otimes SE(3) \quad (16)$$

with the corresponding Lie algebra

$$\mathfrak{se}(3)^n := \mathfrak{se}(3) \oplus \ldots \oplus \mathfrak{se}(3). \quad (17)$$

The Lie exponential and logarithm maps for $(B_1,\ldots,B_n) \in \mathfrak{se}(3)^n$ and $(P_1,\ldots,P_n) \in SE(3)^n$ are given by

$$\exp_{SE(3)^n}(B_1,\ldots,B_n) = (\mathbf{e}^{B_1},\ldots,\mathbf{e}^{B_n}),$$
$$\log_{SE(3)^n}(P_1,\ldots,P_n) = (\mathbf{log}(P_1),\ldots,\mathbf{log}(P_n)). \quad (18)$$

Interpolation on $SE(3)^n$ can be performed by simultaneously interpolating on individual $SE(3)$.

4. Quaternions and dual quaternions

In this section, we provide a brief introduction to quaternions and dual quaternions. While quaternions are commonly used to represent 3D rotations, dual quaternions are used to represent the full 3D rigid body transformations. We refer the readers to Broida et al. (1990); Kavan et al. (2008); Kenwright (2012); McCarthy (1990); Young and Chellappa (1990) for further details on these topics.

4.1. Quaternions

The set of quaternions $\mathcal{Q}$ is equivalent to the 4-dimensional vector space $\mathcal{R}^4$ equipped with the quaternion multiplication operation. Let $\{\mathbf{e_0}, \mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}\}$ be the canonical basis for the vector space $\mathcal{R}^4$. The quaternion multiplication is defined by giving the following multiplication table for the basis:

$$\begin{aligned} &\mathbf{e_0e_1} = \mathbf{e_1e_0} = \mathbf{e_1}, \quad \mathbf{e_1e_2} = -\mathbf{e_2e_1} = \mathbf{e_3}, \\ &\mathbf{e_0e_2} = \mathbf{e_2e_0} = \mathbf{e_2}, \quad \mathbf{e_2e_3} = -\mathbf{e_3e_2} = \mathbf{e_1}, \\ &\mathbf{e_0e_3} = \mathbf{e_3e_0} = \mathbf{e_3}, \quad \mathbf{e_3e_1} = -\mathbf{e_1e_3} = \mathbf{e_2}, \\ &\mathbf{e_0e_0} = \mathbf{e_0}, \quad \mathbf{e_1e_1} = \mathbf{e_2e_2} = \mathbf{e_3e_3} = -\mathbf{1}. \end{aligned} \quad (19)$$

A 4-dimensional quaternion $\vec{q}$ is commonly represented as $(s_q, \vec{v}_q)$, where $s_q \in \mathcal{R}$ is referred to as the scalar or real part and $\vec{v}_q \in \mathcal{R}^3$ is referred to as the vector or imaginary part. Addition of two quaternions $\vec{p} = (s_p, \vec{v}_p)$ and $\vec{q} = (s_q, \vec{v}_q)$ is given by

$$\vec{p} + \vec{q} = (s_p + s_q, \vec{v}_p + \vec{v}_q). \quad (20)$$

Using (19), multiplication of $\vec{p}$ and $\vec{q}$ can be computed as

$$\vec{p}\vec{q} = (s_p s_q - \vec{v}_p \odot \vec{v}_q, s_p\vec{v}_q + s_q\vec{v}_p + \vec{v}_p \times \vec{v}_q), \quad (21)$$

where $\vec{v}_p \odot \vec{v}_q$ and $\vec{v}_p \times \vec{v}_q$ respectively represent the dot product and cross product between $\vec{v}_p$ and $\vec{v}_q$. Note that quaternion multiplication is not commutative.

The conjugate $\bar{\vec{q}}$, the norm $\|\vec{q}\|$, and the exponential $e^{\vec{q}}$ of a quaternion $\vec{q} = (s_q, \vec{v}_q)$ are given by

$$\bar{\vec{q}} = (s_q, -\vec{v}_q), \quad \|\vec{q}\| = \sqrt{\vec{q}\bar{\vec{q}}} = \sqrt{s_q^2 + \|\vec{v}_q\|_2^2},$$
$$e^{\vec{q}} = \left(e^{s_q}\cos(\|\vec{v}_q\|), \quad e^{s_q}\sin(\|\vec{v}_q\|)\frac{\vec{v}_q}{\|\vec{v}_q\|}\right). \quad (22)$$

The quaternions with unit norm are known as *unit quaternions*. The set of unit quaternions, denoted by $\mathcal{UQ}$, forms a Lie group with quaternion multiplication as the group multiplication operation, and $\vec{q}_e = (1, \vec{0})$ as the group identity element. The Lie algebra of $\mathcal{UQ}$, denoted by $\mathfrak{uq}$, is the three dimensional vector space spanned by the set of purely imaginary quaternions. The Lie exponential and logarithm maps for $\vec{w} \in \mathfrak{uq}$ and $\vec{q} = (s_q, \vec{v}_q) \in \mathcal{UQ}$ are given by

$$\exp_{\mathcal{UQ}}(\vec{w}) = e^{\vec{w}} = \left(\cos(\|\vec{w}\|),\ \sin(\|\vec{w}\|)\frac{\vec{w}}{\|\vec{w}\|}\right),$$
$$\log_{\mathcal{UQ}}(\vec{q}) = \cos^{-1}(s_q)\frac{\vec{v}_q}{\sqrt{1-s_q^2}}. \quad (23)$$

Unit quaternions are commonly used to represent rotations in 3D space. Let $\vec{z}$ be a 3D point, and $\vec{q}_z = (0, \vec{z})$ be its quaternion representation. Let $\vec{z}'$ be a 3D point obtained by rotating $\vec{z}$ by an angle θ about an axis $\vec{n}$, and $\vec{q}_{z'} = (0, \vec{z}')$ be its quaternion representation. Then, we have $\vec{q}_{z'} = \vec{r}\vec{q}_z\bar{\vec{r}}$, where $\vec{r} = e^{\vec{n}\frac{\theta}{2}}$. Hence, the unit quaternion

$$e^{\vec{n}\frac{\theta}{2}} = \left(\cos\left(\frac{\theta}{2}\right),\ \vec{n}\sin\left(\frac{\theta}{2}\right)\right) \quad (24)$$

represents the 3D rotation by an angle θ about the axis $\vec{n}$. We can easily convert between unit quaternion and $SO(3)$ representations using

$$\vec{r} = \exp_{\mathcal{UQ}}\left(\frac{\vec{w}}{2}\right), \ \vec{w} = \text{vec}\left(\log_{SO(3)}(R)\right),$$
$$R = \exp_{SO(3)}(skew(\vec{w})), \ \vec{w} = 2\log_{\mathcal{UQ}}(\vec{r}). \quad (25)$$

$\mathcal{UQ} \otimes \ldots \otimes \mathcal{UQ}$: We can combine multiple $\mathcal{UQ}$ groups using the direct product $\otimes$ to form a new Lie group

$$\mathcal{UQ}^n := \mathcal{UQ} \otimes \ldots \otimes \mathcal{UQ} \quad (26)$$

with the corresponding Lie algebra

$$\mathfrak{uq}^n := \mathfrak{uq} \oplus \ldots \oplus \mathfrak{uq}. \quad (27)$$

4.2. Dual quaternions

The set of dual quaternions $\mathcal{D}$ is the extension of quaternions using dual number theory. Each dual quaternion consists of eight elements or two quaternions:

$$\vec{\zeta} = \vec{q}_r + \epsilon\vec{q}_d, \quad (28)$$

where $\vec{q}_r = (s_r, \vec{v}_r), \vec{q}_d = (s_d, \vec{v}_d)$ are quaternions, and ϵ is the dual operator, i.e., $\epsilon^2 = 0, \epsilon \neq 0$. The dual quaternion addition, multiplication, conjugate and magnitude are given by

$$\begin{aligned}
\vec{\zeta}_1 + \vec{\zeta}_2 &= (\vec{q}_{1r} + \vec{q}_{2r}) + \epsilon(\vec{q}_{1d} + \vec{q}_{2d}),\\
\vec{\zeta}_1\vec{\zeta}_2 &= \vec{q}_{1r}\vec{q}_{2r} + \epsilon(\vec{q}_{1r}\vec{q}_{2d} + \vec{q}_{1d}\vec{q}_{2r}),\\
\bar{\vec{\zeta}} &= \bar{\vec{q}}_r + \epsilon\bar{\vec{q}}_d,\\
\|\vec{\zeta}\| &= \sqrt{\vec{\zeta}\bar{\vec{\zeta}}} = \|\vec{q}_r\| + \epsilon\left(\frac{s_r s_d + \vec{v}_r \odot \vec{v}_d}{\|\vec{q}_r\|}\right).
\end{aligned} \quad (29)$$

Note that the magnitude of dual quaternion is a dual number. Dual quaternions that satisfy

$$\|\vec{\zeta}\| = 1, \ \text{i.e.,} \ \|\vec{q}_r\| = 1, \ s_r s_d + \vec{v}_r \odot \vec{v}_d = 0, \quad (30)$$

are called *unit dual quaternions*. We denote the set of all unit dual quaternions using $\mathcal{UD}$.

While a unit quaternion can represent a 3D rotation, a unit dual quaternion can represent a full 3D rigid body transformation, i.e, both rotation and translation. Let $\vec{z}$ be a 3D point, and $\vec{\zeta}_z = (1, \vec{0}) + \epsilon(0, \vec{z})$ be its dual quaternion representation. Let $\vec{z}'$ be a 3D point obtained by transforming $\vec{z}$ using a rotation by an angle θ about an axis $\vec{n}$ followed by a translation $\vec{d}$, and $\vec{\zeta}_{z'} = (1, \vec{0}) + \epsilon(0, \vec{z}')$ be its dual quaternion representation. Then, we have $\vec{\zeta}_{z'} = \vec{\zeta}_{rd}\vec{\zeta}_z\bar{\vec{\zeta}}_{rd}$, where

$$\begin{aligned}
\vec{\zeta}_{rd} &= \vec{r} + \epsilon\left(\frac{1}{2}\vec{t}\vec{r}\right) \in \mathcal{UD},\\
\vec{r} &= e^{\vec{n}\frac{\theta}{2}} \in \mathcal{UQ}, \ \vec{t} = (0, \vec{d}) \in \mathcal{Q}.
\end{aligned} \quad (31)$$

Hence, the unit dual quaternion $\vec{\zeta}_{rd}$ represents the 3D rigid body transformation composed of a rotation by an angle θ about an axis $\vec{n}$ and a translation $\vec{d}$.

5. Relative 3D geometry-based skeletal representations

Let $S = (V, E)$ be a skeleton (Fig. 1), 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. Let e_{m1}, e_{m2} respectively denote the starting and end points of e_m.

Given a pair of body parts e_m and e_n, to describe their relative 3D geometry, we use the rigid body transformations required to take one body part to the position and orientation of the other. A full rigid body transformation T is composed of a rotation by an angle θ about an axis $\vec{n}$ and a translation $\vec{d}$. To measure the rigid body transformation $T_{m,n} = (\theta_{m,n}, \vec{n}_{m,n}, \vec{d}_{m,n})$ required to take e_n to the position and orientation of e_m, we use a local coordinate system attached to e_n (Fig. 4(a)). Similarly, to measure the rigid body transformation $T_{n,m} = (\theta_{n,m}, \vec{n}_{n,m}, \vec{d}_{n,m})$ required to take e_m to the position and orientation of e_n, we use a local coordinate system attached to e_m (Fig. 4(b)). We obtain the local coordinate system of a body part e_m by rotating (with minimum rotation) and translating the global coordinate system such that e_{m1} becomes the origin and e_m coincides with the x-axis.

At first glance it might appear that using only $T_{m,n}$ or $T_{n,m}$ would be sufficient to represent the relative geometry between e_m and e_n. Consider the case in which e_n is rotating about an axis parallel to e_m. Though there is relative motion between the two, $T_{m,n}$ will not change. Similarly, if e_m is rotating about an axis parallel to e_n, then $T_{n,m}$ will not change. So, if we represent the relative geometry using only one of them, the representation will not change under certain kinds of relative motions, which is undesirable. Hence, we use both $T_{m,n}$ and $T_{n,m}$ to represent the relative geometry between e_m and e_n. Note that both $T_{m,n}$ and $T_{n,m}$ 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

$$C(t) = (T_{1,2}(t), \ T_{2,1}(t), \ \ldots, \ T_{M-1,M}(t), \ T_{M,M-1}(t)), \quad (32)$$

where M is the number of body parts. The total number of rigid body transformations used in the skeletal representation is $K = M(M-1)$. Using the proposed representation, a skeletal sequence describing an action can be represented as a curve $\{C(t), t \in [0, T']\}$, and action recognition can be performed by classifying such curves into different action categories.

Note that we are using only the relative measurements $T_{m,n}(t)$ 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 body part e_m can be described using its rigid body transformation T_m with respect to global x-axis (Fig. 4(c)). But, this did not give any improvement, suggesting that the absolute measurements are redundant when the relative measurements are used.

5.1. R3DG features

There are multiple ways to mathematically represent the rigid body transformations in 3D space. In this work, we consider the following four representations: $SE(3)$, $SO(3) \otimes \mathcal{R}^3$, $\mathcal{UQ} \otimes \mathcal{R}^3$, and $\mathcal{UD}$. Using each representation we get a full rigid body transformation-based R3DG feature.

SE(3): Each rigid body transformation $T_{i,j}(t)$ is represented as a member of $SE(3)$ using the 4×4 matrix

$$P_{i,j}(t) = \begin{bmatrix} R_{i,j}(t) & \vec{d}_{i,j}(t) \\ 0 & 1 \end{bmatrix}, \quad (33)$$

where $R_{i,j}(t)$ is the $SO(3)$ representation of 3D rotation $\left(\theta_{i,j}(t), \vec{n}_{i,j}(t)\right)$, and the entire skeleton is represented using

$$\left(P_{1,2}(t), \ P_{2,1}(t), \ldots, P_{M-1,M}(t), P_{M,M-1}(t)\right) \in SE(3)^K. \quad (34)$$

Since $SE(3)^K$ is a curved space, classification of action curves in this space is not an easy task. Standard classification approaches like SVM, which are defined for vector space representations, are not directly applicable to the non-vector space $SE(3)^K$. Also, temporal modeling approaches like Fourier analysis are not applicable to this space. Note that the standard Fourier analysis is defined for functions whose output varies along the real line. Here, the action curve $C(t)$ evolves in the non-Euclidean space $SE(3)^K$ as a function of time, and the standard Fourier analysis is not defined

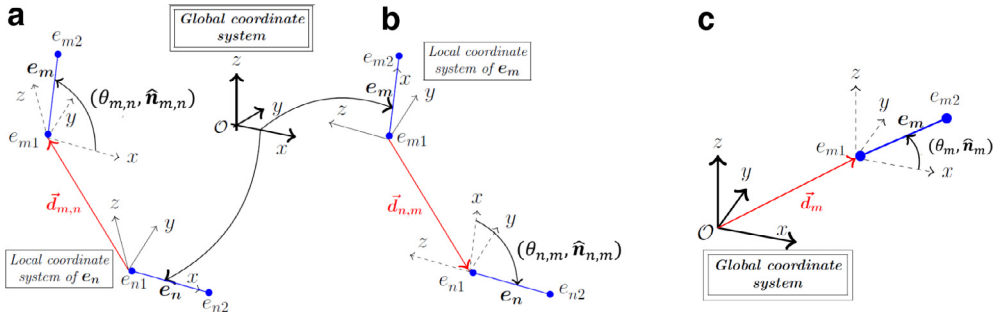

Fig. 4. (a) Rigid body transformation $T_{m,n} = (\theta_{m,n}, \vec{n}_{m,n}, \vec{d}_{m,n})$ from e_n to e_m measured in the coordinate system attached to e_n, (b) Rigid body transformation $T_{n,m} = (\theta_{n,m}, \vec{n}_{n,m}, \vec{d}_{n,m})$ from e_m to e_n measured in the coordinate system attached to e_m, (c) Rigid body transformation $T_m = (\theta_m, \vec{n}_m, \vec{d}_m)$ of e_m with respect to global x-axis.

for this case. To overcome these difficulties, we map the action curves from the Lie group $SE(3)^K$ to its Lie algebra $\mathfrak{se}(3)^K$, which is a $6M(M-1)$-dimensional vector space. The final representation of action curve $C(t)$ is given by

$$\mathfrak{C}_1(t) = \big[\ \mathrm{vec}(\mathbf{log}(P_{1,2}(t))), \mathrm{vec}(\mathbf{log}(P_{2,1}(t))), \ldots, \mathrm{vec}(\mathbf{log}(P_{M-1,M}(t))), \mathrm{vec}(\mathbf{log}(P_{M,M-1}(t)))\big]. \quad (35)$$

$SO(3) \otimes \mathcal{R}^3$: In this case, the rotations and translations are separately represented as members of $SO(3)$ and $\mathcal{R}^3$ respectively, and the entire skeleton is represented using

$$\big(R_{1,2}(t), R_{2,1}(t), \ldots, R_{M-1,M}(t), R_{M,M-1}(t), \vec{d}_{1,2}(t), \vec{d}_{2,1}(t), \ldots, \vec{d}_{M-1,M}(t), \vec{d}_{M,M-1}(t)\big) \in SO(3)^K \otimes \mathcal{R}^{3K}. \quad (36)$$

Similar to $SE(3)^K$, the Lie group $SO(3)^K$ is also a curved space. So, we map the action curves from $SO(3)^K \otimes \mathcal{R}^{3K}$ to the $6M(M-1)$-dimensional vector space $\mathfrak{so}(3)^K \otimes \mathcal{R}^{3K}$ by mapping the rotational part from the Lie group $SO(3)^K$ to its Lie algebra $\mathfrak{so}(3)^K$. Note that the translational part remains the same. The final vector space representation of action curve $C(t)$ is given by

$$\mathfrak{C}_2(t) = \big[\ \mathrm{vec}(\mathbf{log}(R_{1,2}(t))), \mathrm{vec}(\mathbf{log}(R_{2,1}(t))), \ldots, \mathrm{vec}(\mathbf{log}(R_{M-1,M}(t))), \mathrm{vec}(\mathbf{log}(R_{M,M-1}(t))), \vec{d}_{1,2}(t), \vec{d}_{2,1}(t), \ldots, \vec{d}_{M-1,M}(t), \vec{d}_{M,M-1}(t)\big]. \quad (37)$$

$\mathcal{UQ} \otimes \mathcal{R}^3$: In this case, the rotations and translations are separately represented as elements of $\mathcal{UQ}$ and $\mathcal{R}^3$ respectively, and the entire skeleton is represented using

$$\big(\vec{r}_{1,2}(t), \vec{r}_{2,1}(t), \ldots, \vec{r}_{M-1,M}(t), \vec{r}_{M,M-1}(t), \vec{d}_{1,2}(t), \vec{d}_{2,1}(t), \vec{d}_{M-1,M}(t), \vec{d}_{M,M-1}(t)\big) \in \mathcal{UQ}^K \otimes \mathcal{R}^{3K}, \quad (38)$$

where $\vec{r}_{i,j}(t) = (s_{i,j}(t), \vec{v}_{i,j}(t))$ is the unit quaternion representation of 3D rotation $\big(\theta_{i,j}(t), \vec{n}_{i,j}(t)\big)$.

Similar to $SO(3)$ and $SE(3)$, the Lie group $\mathcal{UQ}$ is also a curved surface. In fact, the set of unit quaternions forms a three dimensional unit sphere in $\mathcal{R}^4$. Hence, to get a vector space representation, we directly use the 4-dimensional ambient space representation of unit quaternions. [1] With this, we get the following $7M(M-1)$-dimensional vector space representation for action curve $C(t)$:

$$\mathfrak{C}_3(t) = \big[s_{1,2}(t), \vec{v}_{1,2}(t), s_{2,1}(t), \vec{v}_{2,1}(t), \ldots, s_{M-1,M}(t), \vec{v}_{M-1,M}(t), s_{M,M-1}(t), \vec{v}_{M,M-1}(t), \vec{d}_{1,2}(t), \vec{d}_{2,1}(t), \ldots, \vec{d}_{M-1,M}(t), \vec{d}_{M,M-1}(t)\big]. \quad (39)$$

$\mathcal{UD}$: Each rigid body transformation $T_{i,j}(t)$ is represented using a unit dual quaternion

$$\vec{\zeta}_{i,j}(t) = \big(s^r_{i,j}(t),\ \vec{v}^r_{i,j}(t)\big) + \epsilon\big(s^d_{i,j}(t),\ \vec{v}^d_{i,j}(t)\big). \quad (40)$$

[1] Here, we could have used the Lie algebra representation instead of the ambient space representation. But, the $\mathfrak{uq}$ representation is nothing but a scaled version (a scaling factor of 1/2) of $\mathfrak{so}(3)$ representation (refer to Eq. (25)). Since $\mathfrak{so}(3)$ representation is already being used in the case of $SO(3) \otimes \mathcal{R}^3$, we chose to use the ambient space representation for unit quaternions.

The set of unit dual quaternions does not form a vector space. Hence, similar to the quaternions, we use the 8-dimensional ambient space representation for unit dual quaternions, which gives the following $8M(M-1)$-dimensional vector space representation for action curve $C(t)$:

$$\mathfrak{C}_4(t) = \big[s^r_{1,2}(t), \vec{v}^r_{1,2}(t), s^d_{1,2}(t), \vec{v}^d_{1,2}(t), s^r_{2,1}(t), \vec{v}^r_{2,1}(t), s^d_{2,1}(t), \vec{v}^d_{2,1}(t) \ldots, s^r_{M-1,M}(t), \vec{v}^r_{M-1,M}(t), s^d_{M-1,M}(t), \vec{v}^d_{M-1,M}(t), s^r_{M,M-1}(t), \vec{v}^r_{M,M-1}(t), s^d_{M,M-1}(t), \vec{v}^d_{M,M-1}(t)\big]. \quad (41)$$

5.2. Scale-invariant R3DG features

One of the standard ways to handle scale variations in skeletal data is to resize all the skeletons to a fixed size. This can be done by normalizing the skeletons (without changing the joint angles) such that their body part lengths are equal to the corresponding lengths of a reference skeleton. Interestingly, while the translations between different body parts vary with this scale normalization, the 3D rotations do not change (even if the body parts are not directly connected by a joint). So, by using only the rotations between different body parts, we can get the following two scale-invariant R3DG features based on the $\mathfrak{so}(3)$ and $\mathcal{UQ}$ representations of rotations:

$$\mathfrak{C}_5(t) = \big[\ \mathrm{vec}(\mathbf{log}(R_{1,2}(t))), \mathrm{vec}(\mathbf{log}(R_{2,1}(t))), \ldots, \mathrm{vec}(\mathbf{log}(R_{M-1,M}(t))), \mathrm{vec}(\mathbf{log}(R_{M,M-1}(t)))\big],$$

$$\mathfrak{C}_6(t) = \big[s_{1,2}(t), \vec{v}_{1,2}(t), s_{2,1}(t), \vec{v}_{2,1}(t), \ldots, s_{M-1,M}(t), \vec{v}_{M-1,M}(t), s_{M,M-1}(t), \vec{v}_{M,M-1}(t)\big]. \quad (42)$$

Note that at any time instance t, $\mathfrak{C}_5(t)$ is a $3M(M-1)$-dimensional vector and $\mathfrak{C}_6(t)$ is a $4M(M-1)$-dimensional vector. Table 1 summarizes the proposed family of R3DG features.

6. Temporal modeling and classification

Classification of vector space curves into different action categories is not a straightforward task due to various issues like rate variations, temporal misalignment, noise, etc. Following Kulkarni et al. (2015); Veeraraghavan et al. (2009), we use DTW to handle rate variations. During training, for each action category, we compute a nominal curve using the algorithm described in Table 2, and warp all the training curves to this nominal curve. We use the squared Euclidean distance for DTW computations. Note that for computing a nominal curve all the curves should have equal number of samples. To achieve this, we re-sample the action curves using the interpolation algorithms presented for $SO(3)$ and $SE(3)$ in Section 3.1 and 3.2, respectively. In the case of quaternions, we first interpolate the rotations on $SO(3)$ and then convert them to unit quaternions. In the case of dual quaternions, we first interpolate the rigid body transformations on $SE(3)$ and then convert them to unit dual quaternions.

Table 1
The proposed family of R3DG features.

R3DG feature	Full rigid body transformation-based				Rotation-based	
	$\mathfrak{se}(3)$	$\mathfrak{so}(3) \otimes \mathcal{R}^3$	$\mathcal{UQ} \otimes \mathcal{R}^3$	$\mathcal{UD}$	$\mathfrak{so}(3)$	$\mathcal{UQ}$
Dimensionality	$6M(M-1)$	$6M(M-1)$	$7M(M-1)$	$8M(M-1)$	$3M(M-1)$	$4M(M-1)$
Requires scale normalization	Yes	Yes	Yes	Yes	No	No

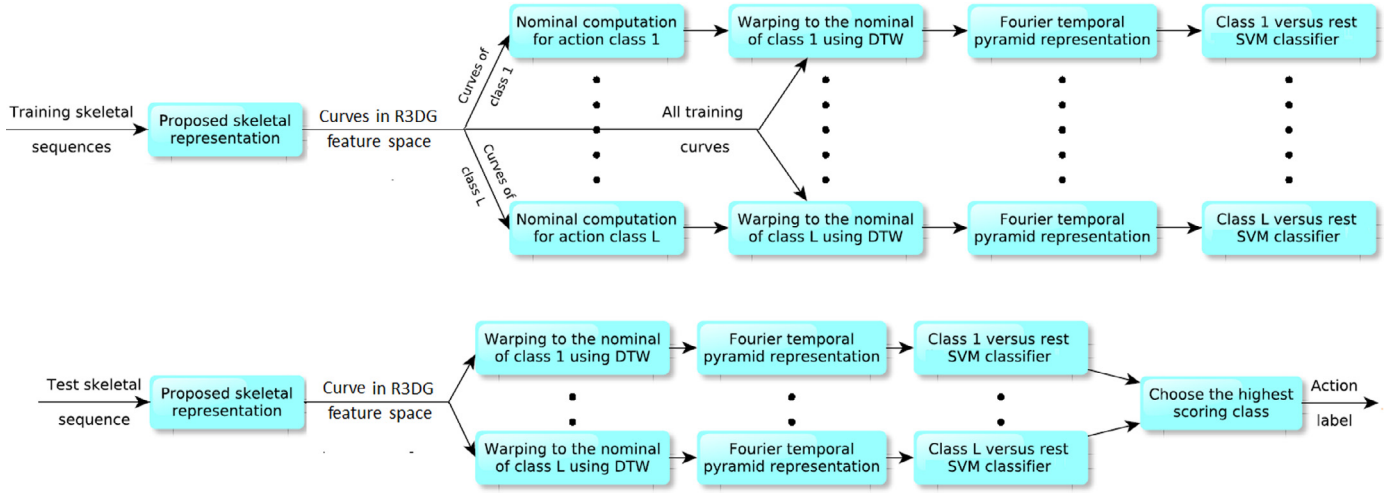

Fig. 5. The top row shows all the steps involved in training and the bottom row shows all the steps involved in testing.

Table 2
Algorithm for computing a nominal curve.

Input: Curves $\mathcal{S}_1(t), \ldots, \mathcal{S}_J(t)$ at $t = 0, 1, \ldots, T'$.
Maximum number of iterations *max* and threshold δ.

Output: Nominal curve $\mathcal{S}(t)$ at $t = 0, 1, \ldots, T'$.

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

After the DTW step, we represent the warped curves using the low-frequency FTP representation that was shown to be robust to temporal misalignment and noise in Wang et al. (2012b). We apply FTP for each dimension separately and concatenate the low-frequency Fourier coefficients to obtain the final feature vector. Action recognition is performed by classifying the final feature vectors using one-vs-all SVM. Fig. 5 gives an overview of the proposed skeleton-based action recognition approach. The top row shows all the steps involved in training and the bottom row shows all the steps involved in testing.

7. Experimental evaluation

In this section, we evaluate the proposed R3DG features on five action datasets: MSRAction3D (Li et al., 2010), UTKinect-Action (Xia et al., 2012), Florence3D (Seidenari et al., 2013), MSRPairs (Oreifej and Liu, 2013) and G3D (Bloom et al., 2012). Please refer to Table 3 for details about these datasets.

Basic pre-processing: To make the skeletal data invariant to the absolute location of human in the scene, all the 3D joint coordinates were transformed from 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 subjects 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. [2] This normalization takes care of scale variations. 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.

Alternative skeletal representations: To show the effectiveness of the proposed R3DG features, we compare them with the following alternative representations:

- **Joint positions (JP):** Concatenation of 3D coordinates of all the joints $v_1, \ldots, v_N$ (except the hip center).
- **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 quaternion representations of all the joint angles. We also tried the $\mathfrak{so}(3)$ and Euler-angle representations for joint angles, but quaternions gave the best results. Here, we measure each joint angle in the local coordinate systems of both body parts associated with that angle.
- **Relation pose features (RP):** We use the joint distance, plane, normal plane, velocity and normal velocity features of Yun et al. (2012) computed from a single human skeleton.
- **Individual body part locations (BPL):** Each body part e_m is represented using its rigid body transformation T_m with respect to global x-axis (Fig. 4(c)). Similar to R3DG features, we have six different BPL features: $\mathfrak{se}(3)$, $\mathfrak{so}(3) \otimes \mathcal{R}^3$, $\mathcal{UQ} \otimes \mathcal{R}^3$, $\mathcal{UD}$, $\mathfrak{so}(3)$, and $\mathcal{UQ}$.

[2] We also performed experiments by varying the reference subject, but the results did not vary much. The standard deviation in the recognition rate was around 0.2–0.3%.

Table 3
Datasets for skeleton-based human action recognition.

MSRAction3D	UTKinect-Action	Florence3D	MSRPairs	G3D
20 actions	10 actions	9 actions	12 actions	20 actions
10 subjects	10 subjects	10 subjects	10 subjects	10 subjects
557 sequences	199 sequences	215 sequences	353 sequences	663 sequences
20 joints	20 joints	15 joints	20 joints	20 joints
19 body parts	19 body parts	14 body parts	19 body parts	19 body parts

Table 4
Alternative skeletal representations for comparison.

Representation	JP	RJP	JA	RP	SQ	
Dimensionality	$3(N-1)$	$\frac{3}{2}N(N-1)$	$8M$	$N(75+\frac{N-1}{2})$	$6M(M-1)$	
Requires scale normalization	Yes	Yes	No	Yes	Yes	
Representation	BPL					
	$\mathfrak{se}(3)$	$\mathfrak{so}(3)\otimes\mathcal{R}^3$	$\mathcal{UQ}\otimes\mathcal{R}^3$	$\mathcal{UD}$	$\mathfrak{so}(3)$	$\mathcal{UQ}$
Dimensionality	$6M$	$6M$	$7M$	$8M$	$3M$	$4M$
Requires scale normalization	Yes	Yes	Yes	Yes	No	No

Table 5
Recognition rates for various skeletal representations on five action datasets.

Representation		MSRAction3D	UTKinect	Florence3D	MSRPairs	G3D
JP		88.75	95.08	85.26	92.90	87.28
RJP		88.87	95.48	85.17	93.91	90.03
JA		75.39	94.07	80.45	90.46	86.25
RP		87.25	93.46	76.86	84.76	88.19
SQ		83.44	95.18	88.89	90.70	89.79
BPL	$\mathfrak{se}(3)$	82.97	94.58	81.38	89.62	87.40
	$\mathfrak{so}(3)\otimes\mathcal{R}^3$	83.88	94.67	81.26	90.59	87.19
	$\mathcal{UQ}\otimes\mathcal{R}^3$	88.74	96.18	84.94	93.32	89.48
	$\mathcal{UD}$	87.76	95.48	83.95	92.75	88.73
	$\mathfrak{so}(3)$	82.46	94.37	80.52	90.70	86.64
	$\mathcal{UQ}$	86.30	95.18	83.46	92.41	87.76
R3DG	$\mathfrak{se}(3)$	89.55	**97.20**	90.71	93.65	91.60
	$\mathfrak{so}(3)\otimes\mathcal{R}^3$	89.37	**97.20**	90.87	93.82	91.60
	$\mathcal{UQ}\otimes\mathcal{R}^3$	90.24	97.09	**92.61**	93.60	91.51
	$\mathcal{UD}$	**90.69**	97.09	91.55	**94.33**	**92.12**
	$\mathfrak{so}(3)$	89.22	96.78	90.52	93.48	91.48
	$\mathcal{UQ}$	90.59	96.88	91.27	93.88	**92.12**

- **Skeletal quads (SQ):** We use the skeletal quad descriptor of Evangelidis et al. (2014) to describe the relative geometry between every pair of body parts.

For a fair comparison, we use the same temporal modeling and classification approach described in Section 6 with all the representations. Table 4 summarizes the alternative skeletal representations used for comparison.

Parameters: For FTP representation, we used a three-level temporal pyramid with one-fourth of the segment length as the number of low-frequency coefficients. While using one or two levels for the temporal pyramid produced inferior results, going beyond three did not improve the results significantly. Changing the number of low-frequency coefficients from one-fourth of the segment length to one-third or one-fifth did not significantly change the accuracy (around 0.2%). The value of SVM parameter C was chosen using cross-validation. For each dataset, all the curves 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.

Comparison with other skeletal representations: Table 5 shows the recognition rates for various skeletal representations on five action datasets when the same temporal modeling and classification pipeline (DTW + FTP + linear SVM) is used. For all the datasets, we followed the cross-subject test setting, in which half of the subjects were used for training and the other half were used for testing. All the results reported in this table were averaged over ten different random combinations of training and test data. The best result in each column is shown in boldface style. We can see that all the proposed R3DG features perform better than all the alternative skeletal representations on all five datasets except the MSR-Pairs dataset where the RJP representation performs slightly better than some of the R3DG features. Specifically, the accuracy of the best R3DG feature is better than the accuracy of the best competing skeletal representation by 1.82% on the MSRAction3D dataset, 1.02% on the UTKinect dataset, 3.72% on the Florence3D dataset, 0.42% on the MSRPairs dataset, and 2.09% on the G3D dataset. These results clearly show the superiority of the proposed R3DG features.

Contribution of the translational information: Comparing the recognition rates of rotation-based and full rigid body transformation-based R3DG features, we can see that on four out of five datasets, the contribution [3] of translational information is not that significant. The difference between the best recognition rates of rotation-based and full rigid body transformation-based R3DG features is 0.1% for the MSRAction3D dataset, 0.32% for the UTKinect dataset, 0.45% for the MSRPairs dataset, and 0% for the

[3] By contribution, we mean the additional contribution of translational information in the presence of rotational information.

Table 6
Contribution of the FTP module in terms of recognition accuracy.

Dataset		$\mathfrak{se}(3)$	$\mathfrak{so}(3) \otimes \mathcal{R}^3$	$\mathcal{UQ} \otimes \mathcal{R}^3$	$\mathcal{UD}$	$\mathfrak{so}(3)$	$\mathcal{UQ}$
MSRAction3D	DTW + SVM	87.71	87.52	90.30	90.59	86.96	90.23
	DTW + FTP + SVM	89.55	89.37	90.24	90.69	89.22	90.59
	FTP contribution	1.84	1.85	-0.06	0.10	2.26	0.36
G3D	DTW + SVM	88.13	88.28	89.73	89.73	87.92	89.52
	DTW + FTP + SVM	91.60	91.60	91.51	92.12	91.48	92.12
	FTP contribution	3.47	3.32	1.78	2.39	3.56	2.60

Table 7
Contribution of the DTW module in terms of recognition accuracy.

Dataset		$\mathfrak{se}(3)$	$\mathfrak{so}(3) \otimes \mathcal{R}^3$	$\mathcal{UQ} \otimes \mathcal{R}^3$	$\mathcal{UD}$	$\mathfrak{so}(3)$	$\mathcal{UQ}$
MSRAction3D	FTP + SVM	86.93	86.94	87.58	87.69	86.42	87.26
	DTW + FTP + SVM	89.55	89.37	90.24	90.69	89.22	90.59
	DTW contribution	2.62	2.43	2.66	3.00	2.80	3.33
G3D	FTP + SVM	91.75	91.75	91.48	92.12	91.60	92.12
	DTW + FTP + SVM	91.60	91.60	91.51	92.12	91.48	92.12
	DTW contribution	–0.15	-0.15	0.03	0	–0.12	0

G3D dataset. Only on the Florence3D dataset, there is a significance difference of around 1.34%.

Contribution of DTW and FTP modules: Our temporal modeling consists of DTW and FTP modules. To analyze the contribution of these modules to the final recognition accuracy, we performed experiments on the MSRAction3D and G3D datasets (the two largest datasets) by removing these modules from our action recognition pipeline. Table 6 compares the final accuracy with and without the FTP module in the action recognition pipeline. As we can see, the FTP module contributes significantly to the final accuracy in most of the cases.

Table 7 compares the final accuracy with and without the DTW module in the action recognition pipeline. While the DTW module contributes significantly to the final accuracy in the case of MSRAction3D dataset, it does not change the accuracy much in the case of G3D dataset. This variation is expected since the contribution of DTW depends on the rate variations present in the dataset.

Comparison with state-of-the-art approaches: Table 8 compares the proposed approach with various existing skeleton-based action recognition approaches on MSRAction3D, UTKinect, Florence3D and G3D datasets. [4] Since the focus of this work is on skeleton-based human action recognition, we use only skeleton-based approaches for comparison. Though combining skeletal features with depth-based features may improve the accuracy, feature fusion is beyond the scope of this work. We evaluated our approach using both linear and RBF kernel SVMs, and the kernel SVM performed slightly better on all datasets. When the RBF kernel was used, the $\mathcal{UQ}$ R3DG feature gave the best result (among all R3DG features) for the G3D dataset and the $\mathcal{UQ} \times \mathcal{R}^3$ R3DG feature gave the best result on all the remaining datasets. For the MSRAction3D dataset, we followed the standard protocol of using subjects 1, 3, 5, 7, 9 for training and the remaining for testing. For G3D, UTKinect and Florence3D datasets, we followed the cross-subject setting of Zhu et al. (2013) and used half of the subjects for training and the remaining for testing. Note that this is a more difficult setting compared to the leave-one-action-out scheme used for the UTKinect dataset in (Devanne et al., 2014; Presti et al., 2014; Xia et al., 2012) and the leave-one-actor-out scheme used for the Florence3D dataset in Devanne et al. (2014); Seidenari et al. (2013), where more subjects were used for training. We report the results averaged over ten random combinations of training and test data.

Table 8
Comparison with the various existing skeleton-based action recognition approaches.

MSRAction3D dataset	
Bag of key poses (Chaaraoui et al., 2013)†	89.62
Random forests (Zhu et al., 2013)†	90.90
HOD features (Gowayyed et al., 2013)†	91.26
Covariance descriptors (Hussein et al., 2013)†	90.53
Spatial and temporal part-sets (Wang et al., 2013)†	90.22
Skeletal quads (Evangelidis et al., 2014)†	89.86
Moving pose (Zanfir et al., 2013)†	91.70
Actionlets (Wang et al., 2012b)	88.20
MMTW (Wang and Wu, 2013)	**92.70**
Motion trajectories (Devanne et al., 2014)	92.10
Hanklets (Presti et al., 2014)	89.00
Joint angle similarities (Ohn-bar and Trivedi, 2013)	83.53
Proposed approach (linear SVM)	89.74
Proposed approach (RBF kernel SVM)	90.11
UTKinect dataset	
Histograms of 3D joints (Xia et al., 2012)**	90.92
Hanklets (Presti et al., 2014)**	86.76
Motion trajectories (Devanne et al., 2014)**	91.50
Random forests (Zhu et al., 2013)	87.90
Elastic functional coding (Anirudh et al., 2015)	94.87
Proposed approach (linear SVM)	97.20
Proposed approach (RBF kernel SVM)	**97.59**
Florence3D dataset	
Multi-part bag-of-poses (Seidenari et al., 2013)*	82.00
Motion trajectories (Devanne et al., 2014)*	87.04
Elastic functional coding (Anirudh et al., 2015)	89.67
Proposed approach (linear SVM)	92.61
Proposed approach (RBF kernel SVM)	**93.06**
G3D dataset	
RBM + HMM (Nie and Ji, 2014)	86.40
Proposed approach (linear SVM)	92.12
Proposed approach (RBF kernel SVM)	**92.39**

† Easier protocol of dividing the dataset into three subsets; * Easier leave-one-actor-out scheme; ** Easier leave-one-action-out scheme.

The best accuracy on each dataset is shown in boldface style. We can see that the proposed approach gives the best classification accuracy on three out of four datasets. Specifically, it outperforms the state-of-the-art results by 2.72% on the UTKinect dataset, 3.39% on the Florence3D dataset and 5.99% on the G3D dataset. The pro-

[4] To the best of our knowledge, all the results reported in the literature for the MSRPairs dataset are based on depth data.

posed approach also outperforms many recent skeleton-based action recognition approaches on the MSRAction3D dataset. Note that the main focus of this work is on skeletal representation, and the proposed R3DG features clearly outperform various existing skeletal representations when the same classification pipeline is used with all the representations.

8. Conclusion and future work

In this paper, we introduced a family of body part-based 3D skeletal representations for human action recognition, which we refer to as R3DG features. The proposed representations explicitly model the relative 3D geometry between various body parts (though not directly connected by a joint) using rigid body transformations. We represented 3D rigid body transformations using $SE(3)$, $SO(3) \otimes \mathcal{R}^3$, $\mathcal{UQ} \otimes \mathcal{R}^3$, and $\mathcal{UD}$, resulting in four different R3DG features. We also introduced scale-invariant R3DG features by using only the 3D rotations between various body parts. Using the proposed representations, we modeled the human actions as curves in R3DG feature spaces. Finally, we performed action recognition by classifying these curves using a combination of dynamic time warping, Fourier temporal pyramid representation and SVM. We experimentally showed that the proposed R3DG features perform better than various existing skeletal representations, and the proposed action recognition approach outperforms various state-of-the-art skeleton-based approaches.

In our work, we used the relative geometry between all pairs of body parts. However, 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. While we focused only on actions performed by a single person in this work, we are planning to extend these representations to model multi-person interactions. We also plan to improve the results further by using the proposed skeletal representations with more complex classification approaches based on sparse representation, deep networks, etc.

Acknowledgements

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

References

Aggarwal, J.K., Ryoo, M.S., 2011. Human activity analysis: a review. ACM Comput. Surv. 43 (3), 16:1–16:43.
Anirudh, R., Turaga, P., Su, J., Srivastava, A., 2015. Elastic functional coding of human actions: from vector-fields to latent variables. In: IEEE Conference on Computer Vision and Pattern Recognition.
Belta, C., Kumar, V., 2002. An SVD-based projection method for interpolation on SE(3). IEEE Trans. Rob. Autom. 18 (3), 334–345.
Bloom, V., Makris, D., Argyriou, V., 2012. G3D: a gaming action dataset and real time action recognition evaluation framework. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.
Broida, T.J., Chandrashekhar, S., Chellappa, R., 1990. Recursive 3-D motion estimation from a monocular image sequence. IEEE Trans. Aerosp. Electron. Syst. 26, 639–656.
Campbell, L.W., Bobick, A.F., 1995. Recognition of human body motion using phase space constraints. In: IEEE International Conference on Computer Vision.
Chaaraoui, A.A., Padilla-López, J.R., Flórez-Revuelta, F., 2013. Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices. In: IEEE International Conference on Computer Vision Workshops.
Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R., Vidal, R., 2013. Bio-inspired dynamic 3D discriminative skeletal features for human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.
Chen, C., Jafari, R., Kehtarnavaz, N., 2015. Action recognition from depth sequences using depth motion maps-based local binary patterns. In: IEEE Winter Conference on Applications of Computer Vision.
Chen, G., Giuliani, M., Clarke, D., Gaschler, A., Knoll, A., 2014. Action recognition using ensemble weighted multi-instance learning. In: IEEE International Conference on Robotics and Automation.
Chen, L., Wei, H., Ferryman, J.M., 2013. A survey of human motion analysis using depth imagery. Pattern Recognit. Lett. 34 (15), 1995–2006.
Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., Bimbo, A.D., 2014. 3D human action recognition by shape analysis of motion trajectories on Riemannian manifold. IEEE Trans. Cybern. 45 (7), 1340–1352.
Du, Y., Wang, W., Wang, L., 2015. Hierarchical recurrent neural network for skeleton based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition.
Ellis, C., Masood, S.Z., Tappen, M.F., LaViola, J.J., Sukthankar, R., 2013. Exploring the trade-off between accuracy and observational latency in action recognition. Int. J. Comput. Vis. 101 (3), 420–436.
Evangelidis, G., Singh, G., Horaud, R., 2014. Skeletal quads: human action recognition using joint quadruples. In: International Conference on Pattern Recognition.
Gavrila, D.M., Davis, L.S., 1995. Towards 3-D model-based tracking and recognition of human movement: a multi-view approach. International Workshop on Automatic Face and Gesture Recognition.
Gowayyed, M.A., Torki, M., Hussein, M.E., El-Saban, M., 2013. Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition. In: International Joint Conference on Artificial Intelligence.
Gu, J., Ding, X., Wang, S., Wu, Y., 2010. Action and gait recognition from recovered 3-D human joints. IEEE Trans. Syst Man Cybern Part B 40 (4), 1021–1033.
Hall, B., 2003. Lie Groups, Lie Algebras, and Representations: An Elementary Introduction. Springer.
Hu, J.-F., Zheng, W.-S., Lai, J., Zhang, J., 2015. Jointly learning heterogeneous features for RGB-D activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition.
Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M., 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In: International Joint Conference on Artificial Intelligence.
Jaakkola, T.S., Haussler, D., 1998. Exploiting generative models in discriminative classifiers. Neural Information Processing Systems.
Johansson, G., 1973. Visual perception of biological motion and a model for its analysis. Percept. Psychophys. 14 (2), 201–211.
Kavan, L., Collins, S., Zára, J., O'Sullivan, C., 2008. Geometric skinning with approximate dual quaternion blending. ACM Trans. Graphics 27 (4), 105:1–105:23.
Kenwright, B., 2012. A Beginners Guide to Dual-Quaternions: What They are, how they work, and how to use them for 3D character hierarchies. In: International Conference on Computer Graphics, Visualization and Computer Vision.
Kong, Y., Fu, Y., 2015. Bilinear Heterogeneous Information Machine for RGB-D Action Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition.
Kulkarni, K., Evangelidis, G., Cech, J., Horaud, R., 2015. Continuous Action Recognition Based on Sequence Alignment. Int. J. Comput. Vis. 112 (1), 90–114.
Li, W., Zhang, Z., Liu, Z., 2010. Action Recognition Based on a Bag of 3D Points. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.
Lu, C., Jia, J., Tang, C., 2014. Range-Sample Depth Feature for Action Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition.
Lv, F., Nevatia, R., 2006. Recognition and Segmentation of 3D Human Action Using HMM and Multi-class AdaBoost. ECCV.
Marr, D., Nishihara, H.K., 1978. Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London, Series B, Biological Sciences 200 (1140), 269–294.
McCarthy, J.M., 1990. Introduction to Theoretical Kinematics. MIT Press.
Moeslund, T.B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. 104 (2-3), 90–126.
Müller, M., 2007. Information Retrieval for Music and Motion. Springer-Verlag, New York.
Müller, M., Röder, T., Clausen, M., 2005. Efficient content-based retrieval of motion capture data. ACM Trans. Graphics 24 (3), 677–685.
Murray, R.M., Li, Z., Sastry, S.S., 1994. A Mathematical Introduction to Robotic Manipulation. CRC press.
Nie, S., Ji, Q., 2014. Capturing global and local dynamics for human action recognition. In: International Conference on Pattern Recognition.
Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R., 2014. Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition. J. Vis. Commun. Image Represent. 25 (1), 24–38.
Ohn-bar, E., Trivedi, M.M., 2013. Joint angles similarities and HOG^2 for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.
Oreifej, O., Liu, Z., 2013. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In: IEEE Conference on Computer Vision and Pattern Recognition.
Park, F.C., Ravani, B., 1997. Smooth invariant interpolation of rotations. ACM Trans. Graphics 16 (3), 277–295.
Presti, L.L., Cascia, M.L., S., S., Camps, O.I., 2014. Gesture modeling by hanklet-based hidden Markov model. In: Asian Conference on Computer Vision.
Rahmani, H., Mahmood, A., Huynh, D.Q., Mian, A.S., 2014. HOPC: histogram of oriented principal components of 3D pointclouds for action recognition. In: European Conference on Computer Vision.
Reyes, M., Dominguez, G., Escalera, S., 2011. Feature weighting in dynamic time warping for gesture recognition in depth data. In: IEEE International Conference on Computer Vision Workshops.
Seidenari, L., Varano, V., Berretti, S., Bimbo, A.D., Pala, P., 2013. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Shao, Z., Li, Y.F., 2013. A new descriptor for multiple 3D motion trajectories recognition. In: IEEE International Conference on Robotics and Automation.
Sheikh, Y., Sheikh, M., Shah, M., 2005. Exploring the space of a human action. In: IEEE International Conference on Computer Vision.
Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-time human pose recognition in parts from a single depth image. In: IEEE Conference on Computer Vision and Pattern Recognition.
Sung, J., Ponce, C., Selman, B., Saxena, A., 2012. Unstructured human activity detection from RGBD images. In: IEEE International Conference on Robotics and Automation.
Turaga, P.K., Chellappa, R., Subrahmanian, V.S., Udrea, O., 2008. Machine recognition of human activities: a survey. IEEE Trans. Circ. Syst. Vid. Technol. 18 (11), 1473–1488.
Veeraraghavan, A., Srivastava, A., Roy-Chowdhury, A.K., Chellappa, R., 2009. Rate-invariant recognition of humans and their activities. IEEE Trans. Image Process. 18 (6), 1326–1339.
Vemulapalli, R., Arrate, F., Chellappa, R., 2014. Human action recognition by representing 3D skeletons as points in a lie group. In: IEEE Conference on Computer Vision and Pattern Recognition.
Wang, C., Wang, Y., Yuille, A.L., 2013. An approach to pose-based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition.
Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y., 2012. Robust 3D action recognition with random occupancy patterns. In: European Conference on Computer Vision.
Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras. In: IEEE Conference on Computer Vision and Pattern Recognition.
Wang, J., Wu, Y., 2013. learning maximum margin temporal warping for action recognition. In: IEEE International Conference on Computer Vision.
Wei, P., Zheng, N., Zhao, Y., Zhu, S.C., 2013. Concurrent action detection with structural prediction. In: IEEE International Conference on Computer Vision.
Xia, L., Aggarwal, J.K., 2013. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: IEEE Conference on Computer Vision and Pattern Recognition.
Xia, L., Chen, C.C., Aggarwal, J.K., 2012. View invariant human action recognition using histograms of 3D joints. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.
Yacoob, Y., Black, M.J., 1998. Parameterized modeling and recognition of activities. In: IEEE International Conference on Computer Vision.
Yang, X., Tian, Y., 2014. Effective 3D action recognition using Eigen joints. J. Vis. Commun. Image Represent. 25 (1), 2–11.
Yang, X., Tian, Y., 2014. Super normal vector for activity recognition using depth sequences. In: IEEE Conference on Computer Vision and Pattern Recognition.
Yang, X., Zhang, C., Tian, Y., 2012. Recognizing actions using depth motion maps-based histograms of oriented gradients. In: ACM Multimedia Conference.
Yao, A., Gall, J., Fanelli, G., Gool, L.V., 2011. Does human action recognition benefit from pose estimation? In: British Machine Vision Conference.
Yao, A., Gall, J., Gool, L.V., 2012. Coupled action recognition and pose estimation from multiple views. Int. J. Comput. Vis. 100 (1), 16–37.
Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., Gall, J., 2013. A survey on human motion analysis from depth data. In: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications. Springer, pp. 149–187.
Young, G.S.J., Chellappa, R., 1990. 3D motion estimation using a sequence of noisy stereo images: models, estimation, and uniqueness results. IEEE Trans. Pattern Anal. Mach. Intell. 12 (8), 735–759.
Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D., 2012. Two-person interaction detection using body-pose features and multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.
Zanfir, M., Leordeanu, M., Sminchisescu, C., 2013. The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection. In: IEEE International Conference on Computer Vision.
Zatsiorsky, V.M., 1997. Kinematics of Human Motion. Human Kinetics Publishers.
Zefran, M., Kumar, V., 1998. Two methods for interpolating rigid body motions. In: IEEE International Conference on Robotics and Automation.
Zefran, M., Kumar, V., Croke, C., 1996. Choice of Riemannian metrics for rigid body kinematics. In: ASME Design Engineering Technical Conference and Computers in Engineering Conference.
Zhang, C., Tian, Y., 2015. Histogram of 3D facets: A depth descriptor for human action and hand gesture recognition. Comput. Vis. Image Underst. 139, 29–39.
Zhu, Y., Chen, W., Guo, G., 2013. Fusing spatiotemporal features and joints for 3D action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.
Zhu, Y., Chen, W., Guo, G., 2014. Evaluating spatiotemporal interest point features for depth-based action recognition. Image Vis. Comput. 32 (8), 453–464.