RRV: A Spatiotemporal Descriptor for Rigid Body Motion Recognition

Yao Guo, Student Member, IEEE, Youfu Li, Senior Member, IEEE, and Zhanpeng Shao

Abstract—Motion behaviors of a rigid body can be characterized by a 6-dimensional motion trajectory, which contains position vectors of a reference point on the rigid body and rotations of this rigid body over time. This paper devises a Rotation and Relative Velocity (RRV) descriptor by exploring the local translational and rotational invariants of motion trajectories of rigid bodies, which is insensitive to noise, invariant to rigid transformation and scaling. A flexible metric is also introduced to measure the distance between two RRV descriptors. The RRV descriptor is then applied to characterize motions of a human body skeleton modeled as articulated interconnections of multiple rigid bodies. To illustrate the descriptive ability of the RRV descriptor, we explore it for different rigid body motion recognition tasks. The experimental results on benchmark datasets demonstrate that this simple RRV descriptor outperforms the previous ones regarding recognition accuracy without increasing computational cost.

Index Terms—Rigid body motion trajectory; translational and rotational invariants; RRV descriptor; motion recognition

I. INTRODUCTION

To realize effective robot-human interaction, robots are expected to show the potential for recognizing motion behaviors of objects or individuals from visual observations. Motion trajectories produced by objects and individuals can provide abundant clues on mining spatiotemporal information in motion characterization [1]. In recent years, a number of 3D trajectory-based gait recognition [2], gesture recognition [3], sign language recognition [4] and human action recognition [5] have been proposed. The significant inherent challenges in trajectory-based recognition are undesirable variations of raw data mainly induced by the noise contamination, viewpoint changing and intra-class variations performed by different individuals. Instead, an invariant descriptor can offer advantages over the raw data in capturing spatiotemporal features of motion behaviors [4]. Most of the previous descriptors are only proposed for 3D point trajectories matching and recognition [4, 6-8]. However, motion behaviors of an object cannot be fully characterized by a 3D trajectory of a reference point on the object, since this point trajectory is insufficient for description without considering rotations of the object.

To this end, the motion trajectory of a rigid body will be taken into account in this paper, and it can be described by the position vectors of a reference point on the rigid body and the rotations of this rigid body across time instances [9]. It is necessary to propose a new corresponding descriptor for representing 6D trajectories. Rigid body motion recognition has attracted much attention in recent years, where the instantaneous screw axis (ISA) descriptor [10, 11] and the SoSaLe descriptor [12] have achieved some successes in their applications. The ISA descriptor is a 6D representation based on a motion model for the ISA. Two of six invariants are the translational speed and the rotational speed along the ISA, and the other four invariants can be derived from the first-order and second-order kinematics of ISA. The SoSaLe descriptor establishes two frames to describe the variation of the position and orientation of a rigid body. These two local frames are constructed according to the Frenet-Serret formulas, where the second-order time derivatives of the translational and rotational velocities are involved [13]. The SoSaLe descriptor is also a 6D representation, where two invariants are the translational and rotational speeds, and the other four are angular variations between two adjacent frames. However, some inherent challenges are unresolved in these two descriptors. Firstly, they both involve high-order time derivatives of raw discrete trajectories, which makes them sensitive to noise and outliers. In addition, the local spatiotemporal relationship between translational and rotational invariants has not been taken into consideration. Moreover, they only propose the descriptors for a single trajectory, the applications on multiple rigid bodies are not investigated.

Motivated by these challenges, this paper firstly proposes a Rotation and Relative Velocity (RRV) descriptor to represent a rigid body motion trajectory. Our RRV descriptor, which captures the local relationship between translational and rotational features of the trajectory, only involves first-order time derivatives of the raw data. Specifically, the RRV descriptor consists of a 4D rotational invariant part and a 3D translational invariant part. For the rotational invariant part, the unit quaternion is chosen to describe 3D rotations of the rigid body, where the rotation vectors are firstly re-parameterized by the singular value decomposition (SVD) method. Meanwhile, the square root velocity function (SRVF) [14, 15] will be selected as the translational invariant part by being projected onto the instantaneous re-parameterized rotation matrix, which can further explore the correlations between point trajectory and rigid body’s pose. In addition, a flexible metric is introduced to measure the distance between two RRV descriptors. We then extend the RRV descriptor to represent human actions. The skeleton of a human body can be viewed as articulated interconnections of multiple rigid bodies. For integrating our RRV
A novel virtual rigid body (VRB) method is proposed to represent rigid body motion trajectories. Such descriptor explores the local rotational and translational features, which can improve the intra-class similarity and inter-class discriminability than the previous descriptors.

The RRV descriptor is extended as a representation for multiple trajectories of human actions. To our knowledge, this is the first application of descriptors incorporating the rotational invariants on the motion recognition for multiple rigid bodies.

A novel virtual rigid body (VRB) method is proposed to provide a discriminative representation for multiple rigid bodies in each hierarchical component of a skeleton, thus improving recognition performance.

We demonstrate that our proposed descriptor is an efficient and discriminative motion representation for motion recognition tasks by conducting extensive experiments on two popular benchmark datasets.

This paper is organized as follows. In Section II, we derive the RRV descriptor by exploring rotational and translational invariants of a rigid body motion trajectory. In Section III, we discuss the properties of the RRV descriptor. A flexible metric is also defined to measure the distance between two RRV descriptors. This RRV descriptor is applied for characterizing the skeleton of a human body in the form of multiple rigid bodies in Section IV. The recognition approaches are addressed in Section V. The experimental studies on two benchmark datasets are conducted to demonstrate the effectiveness and consistency of the RRV descriptor in Section VI.

II. RRV DESCRIPTOR

A. Rigid body motion trajectory

Motion behaviors of a rigid body can be characterized by a 6D trajectory [9] as shown in fig. 2, which consists of the position vectors of a reference point on the rigid body and the rotations with respect to rigid body coordinate system over time, where the reference point is set as the origin. The trajectory parameterized by time index \( t \) can be written as:

\[
\mathbf{m}(t) = [x(t), y(t), z(t), \psi(t), \phi(t), \theta(t)]^T
\]

where \([x, y, z]^T = [x(t), y(t), z(t)]^T\) is the position vector of the reference point recorded in the global coordinate system \( \{E\} \). \( \psi(t), \phi(t) \) and \( \theta(t) \) are Euler angles with respect to the rigid body coordinate system \( \mathbf{B}(t) = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\} \), where \( \mathbf{b}_1, \mathbf{b}_2, \) and \( \mathbf{b}_3 \) are three principal axes. Mathematically, the rotations along \( x, y \) and \( z \)-axes can be called as roll, pitch and yaw respectively.

This paper aims to propose an invariant descriptor \( \mathbf{S} \), which can offer sufficient descriptive power and be invariant to undesirable variations over the raw data. Our descriptor is expected to capture the local rotational and translational invariants of rigid body motion trajectories. It should be noted that rotations and positions are measured in different scales. Therefore, this new descriptor \( \mathbf{S} \) will be divided into the following two parts: rotational invariant part \( \mathbf{S}_r \) and translational invariant part \( \mathbf{S}_t \).

B. Rotational invariants

Euler angles and rotation matrix are two popular ways of representing rotations in 3D. However, these two expressions have some limitations. With Euler angles, the 3D rotation needs...
to be realized in a specified order. The rotation matrix will produce the round-off errors after several matrix multiplications, which makes the matrix non-orthogonal. Moreover, a rotation matrix needs to store more data.

Unit quaternion, also known as versors, offers a convenient notation for representing 3D rotations. According to Euler’s rotation theorem, any rotation in 3D space is equivalent to a rotation by an angle $\beta$ with respect to a fixed Euler axis $\mathbf{w} = \beta \hat{\mathbf{w}}$ (this axis is also called as the rotation vector) [19]. $\hat{\mathbf{w}}$ is a unit vector denoting the direction of rotation and can be derived from the rotation matrix $\mathbf{R}$:

$$\hat{\mathbf{w}} = \frac{1}{2 \sin \beta} \begin{bmatrix} \mathbf{R}(3,2) - \mathbf{R}(2,3) \\ \mathbf{R}(1,3) - \mathbf{R}(3,1) \\ \mathbf{R}(2,1) - \mathbf{R}(1,2) \end{bmatrix}$$  \hspace{1cm} (2)

and $\beta = \arccos(\text{trace}(\mathbf{R}) - 1)/2$ denotes the magnitude of rotation with respect to the axis $\mathbf{w}$. Then the unit quaternion can be written as:

$$\mathbf{q} = \pm \left[ \cos \left( \frac{\beta}{2} \right), \mathbf{w}^T \sin \left( \frac{\beta}{2} \right) \right]^T$$  \hspace{1cm} (3)

In contrast to Euler angles, the unit quaternion is simpler to integrate and can avoid gimbal lock. Compared to the rotation matrix, it is more numerically stable and efficient [20].

However, the expressions for representing 3D rotations (Euler angles, rotation matrix and unit quaternion) will vary when the motions of rigid bodies are observed by a rotated vision system. Hence, the above expressions for 3D rotation will produce confusions under rotational variations.

Without loss of generality, we can assume the rotation matrix of a rotated vision system as $\Gamma$, then the rotated global coordinate system is $\{GE\}$. Mathematically, $\Gamma$ belongs to the special orthogonal group $\text{SO}(n)$, where $\Gamma^T \Gamma = \Gamma \Gamma^T = \mathbf{I}$ and $\det(\Gamma) = 1$. The angle $\beta$ remains unchanged under rotational variations. However, the Euler axis $\mathbf{w}$ in the original coordinate system $\{E\}$ differs from that captured in the $\{GE\}$. To remove the uncertainties induced by rotational variations, the singular value decomposition (SVD) approach will be firstly adopted. Here we denote the matrix $\mathbf{A} = \{\hat{\mathbf{w}}(t), \ldots, \hat{\mathbf{w}}(t_v)\}$ as sequences of the unit direction vectors $\hat{\mathbf{w}}$ for all time instances, where $\mathbf{A} \in \mathbb{R}^{3 \times N}$ is the field of real numbers. After applying SVD, $\mathbf{A}$ can be decomposed as:

$$\mathbf{A} = \mathbf{U}(\Sigma \mathbf{V})^T$$  \hspace{1cm} (4)

where $\mathbf{U}$ is a $3 \times 3$ orthogonal matrix, $\Sigma$ has the same dimensions as $\mathbf{A}$, which is a $3 \times N$ rectangular matrix with non-negative real numbers on the diagonal. $\mathbf{V}^T$ is the transpose of an $N \times N$ matrix $\mathbf{V}$. As well known, the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{A}^T \mathbf{A}$. The entries on the diagonal of $\Sigma$ are called as singular values, which are the square roots of the eigenvalues of $\mathbf{A}^T \mathbf{A}$. In addition, we have:

$$(\Gamma \mathbf{A})^T(\Gamma \mathbf{A}) = \mathbf{A}^T \mathbf{A}$$  \hspace{1cm} (5)

It can be observed that $\mathbf{A}^T \mathbf{A}$ is invariant to rotational variations. In other words, $\mathbf{A}^T \mathbf{A}$ and $(\Gamma \mathbf{A})^T(\Gamma \mathbf{A})$ share the same eigenvectors and eigenvalues. Consequently, the matrices $\Sigma$ and $\mathbf{V}^T$ will remain unchanged under rotational variations. Without loss of generality, $\mathbf{U}$ can be viewed as a rotation matrix. Then we can re-parameterize $\mathbf{A}$ by pre-multiplying it with $\mathbf{U}^T$ as:

$$\hat{\mathbf{A}} = \mathbf{U}^T \mathbf{A} = \Sigma \mathbf{V}^T = \{\hat{\mathbf{w}}(t), \ldots, \hat{\mathbf{w}}(t_v)\}$$  \hspace{1cm} (6)

where $\hat{\mathbf{w}}(t) = \mathbf{U}^T \hat{\mathbf{w}}(t)$ is the re-parameterization of the unit direction vector $\hat{\mathbf{w}}$.

In conclusion, the unit vector $\hat{\mathbf{w}}$ is an invariant representation under rotational variations. Then the re-parameterized unit quaternion $\hat{\mathbf{q}}$ can be calculated by (3). The rotational invariants in our RRV descriptor can be written as $\mathbf{S} = \hat{\mathbf{q}}$.

C. Translational invariants

The 3D point trajectory is a set of position vectors of a reference point on the rigid body, which can be parameterized by the discrete time index $t$:

$$[x(t), y(t), z(t)]^T$$  \hspace{1cm} (7)

where $t \in [1, N]$. However, raw trajectories are under rigid transformation and scaling in $\mathbb{R}^3$ while viewed from different perspectives. Moreover, the variability in the shape and the execution rates usually happen when the same motion trajectory is performed by different objects or individuals. Hence, the effective translational invariant part $\mathbf{S}_t$ for describing 3D point trajectories can offer substantial advantages over raw data.

Square root velocity function (SRVF) has been proposed for analyzing the shapes of 3D trajectories with unit length [14, 15]. This representation only requires first order time derivatives and has shown discriminative power in the description. The SRVF of a 3D point trajectory $\gamma(s)$ is given as:

$$\beta(s) = \left\| \gamma(s) \right\|_I \hspace{1cm} (8)$$

where $s \in [0,1]$ is the discrete unit arc-length and $\left\| \right\|$ is $L_2$-norm. The proposed SRVF descriptor shows richness in describing the local shape information of point trajectories, and it is invariant to the rate variance. However, the directions of velocity vectors vary significantly under rotational variations, thus resulting in unstable descriptive ability.

To this end, the translational invariant part $\mathbf{S}_t$ in RRV descriptor extends the conventional SRVF to rotational invariance. In addition, it further explores the relationship between point trajectory and the rigid body’s pose by means of representing SRVF's in a local coordinate system. Commonly, the local coordinate system is the fixed one attached to the rigid body. However, the three principal axes of the rigid body cannot be consecutive extracted from different perspectives due to occlusions. In this paper, the instantaneous rotation matrix can be alternatively selected as the local coordinate system. Specifically, the re-parameterized rotation matrix $\mathbf{R} = [\mathbf{r}, \tilde{\mathbf{r}}, \bar{\mathbf{r}}]$ can be derived from the re-parameterized unit quaternion $\hat{\mathbf{q}} = [q_x, q_y, q_z, q_d]$:

$$\tilde{\mathbf{r}} = \begin{bmatrix} 1 - 2q_d^2 - 2q_d^2 & 2q_dq_y - 2q_dq_x & 2q_dq_z + 2q_dq_w \\ 2q_dq_y + 2q_dq_x & 1 - 2q_d^2 - 2q_d^2 & 2q_dq_z - 2q_dq_w \\ 2q_dq_z - 2q_dq_w & 2q_dq_w + 2q_dq_z & 1 - 2q_d^2 - 2q_d^2 \end{bmatrix}$$  \hspace{1cm} (9)

Compared to frames established according to Frenet-Serret formulas in [12], this local coordinate system $\tilde{\mathbf{r}}$ only involves first-order time derivatives, which is more robust to noise and outliers. In addition, this system indicates the instantaneous rotational status of the rigid body.
Accordingly, the point trajectory $\gamma(t)$ will be rotated as $\tilde{\gamma}(t) = U^T \gamma(t)$, where $U^T$ is introduced in (4). Let denote $v_\xi$ is the velocity of $\tilde{\gamma}(t)$ and the length of $\tilde{\gamma}(t)$ is normalized to 1. In order to capture the local relationship between the translational and rotational features, $v_\xi$ will be projected onto three principal axes $r_x$, $r_y$, and $r_z$ as shown in Fig. 3, which can be written as

$$v_\xi = \tilde{R}^T v_\xi = \left[ \begin{array}{c} [\tilde{u}_1] v_\xi, [\tilde{u}_2] v_\xi, [\tilde{u}_3] v_\xi \end{array} \right]^T$$  (10)

where $[\tilde{u}_i] v_\xi$ is the velocity of the point trajectory expressed in the local coordinate system $\{\tilde{R}\}$. At last, the SRVF of the projected velocity $[\tilde{u}_i] v_\xi$ is selected as the translational invariant part of our descriptor, which is:

$$S_t = \frac{v_\xi}{\sqrt{\|v_\xi\|}}$$  (11)

This simple projection suffices to capture the local spatio-temporal information of the rigid body motion trajectory, thus improving the descriptiveness power.

D. RRV descriptor

The invariant descriptor for representing rigid body motion trajectories can be called as the Rotation and Relative Velocity (RRV) descriptor. It has two parts to represent the rotational and translational invariants respectively, and can be written as:

$$S = [S_r, S_t]^T$$  (12)

where $S_r \in \mathbb{R}^{4 \times N}$ is the rotational invariant part represented by the re-parameterized unit quaternion, and $S_t \in \mathbb{R}^{3 \times N}$ is the translational invariant part that calculates the SRVF of the point trajectory in a local coordinate system.

III. INVARINANTS AND PROPERTIES

A. Invariant properties

As we claim, the RRV descriptor can show strong invariance to different transformations in 3D space. In the following, the invariant properties of the RRV descriptor will be declared.

1) Invariant to translation

3D rotations and translational velocities are first-order time derivatives, and they will remain unchanged with respect to the translation in 3D space. Hence, the RRV descriptor $S$ is invariant to translation.

2) Invariant to scaling

Scaling can be simplified as the zoom in and zoom out of 3D point trajectories in visual observations. In our method, the length of point trajectories will be normalized as 1 to remove scaling influences.

3) Invariant to rotation

Rotational variations of rigid body motion trajectories are mainly attributed to different viewpoints of vision systems. As presented in Section II, the matrix $A = [w(t_1), \ldots, w(t_N)]$ represents a sequence of directions of 3D rotations. Then we have $A = U \Sigma V^T$ by applying SVD. The rotational invariant part in the RRV descriptor is re-parameterized by $U^T$, which is invariant to rotation. Similarly, the point trajectory also can be rotated by $U^T$ as $\tilde{\gamma}(t)$ to remove rotational variations. Then $S_r$ calculates the SRVF of $\tilde{\gamma}(t)$ in a local coordinate system, which is also invariant to rotation.

4) Invariant to occlusion

Occlusion is a common challenge in motion matching and recognition. Until now, the occlusion cannot be fully handled by global descriptors. On the contrary, our descriptor is capable of handling occlusion naturally as it is with computational locality. The partial descriptions for the non-occluded portion of a full motion sequence can still be generated invariantly. Thereby, with our descriptors the recognition can still be performed under partial occlusion.

B. Metric between RRV descriptors

An appropriate metric to calculate the distance between two RRV descriptors will improve the matching and recognition performance. The previous works [10-12] only calculate the $l_2$-norm between two multivariate descriptors, which can be insufficient while components in the descriptor are represented in different scales. In this paper, a flexible metric is proposed for calculating distance between two RRV descriptors. Recall that the RRV descriptor has the rotational invariant part and the translational invariant part. Inspired by this, the robust metric between these two descriptors $S^p$ and $S^q$ can be defined as:

$$d(p, q) = \min \left\{ \|S_r^p - S_r^q\|, \|S_t^p + S_t^q\|, \|S_t^p - S_t^q\| \right\}$$  (13)

where the first part denotes the distance between two unit quaternions [20], and the second one is the distance between two local SRVFs.

C. Special patterns

We then discuss the RRV descriptor for some special patterns in this sub-section.

1) Rotation with incomplete observations

In some special situations, for instance, the rigid body is the cylinder or the conic, three principal axes $b_x$, $b_y$, and $b_z$ cannot be extracted simultaneously from visual observations. To solve this, 3D rotations of the rigid body can be alternatively calculated by an axis $b_x$, where rotations along $b_x$ will not be taken into consideration. Given two adjacent principal axes $b_x(t)$ and $b_x(t+1)$, the rotation matrix $R(t)$ can be derived from the following formulas. Let denote $a_1 = b_x(t)/\|b_x(t)\|$ and $a_2 = b_x(t+1)/\|b_x(t+1)\|$ as two unit vectors. The cross-product of these two vectors is $c = a_1 \times a_2$. Then $R(t)$ can be calculated by:

![Fig. 3. The sketch of the projected velocity. The velocity vector of the reference point will be projected onto the three axes of re-parameterized rotation matrix $R$.](image)
\[ \mathbf{R}(t) = \mathbf{I} + [\mathbf{c}]_\times + [\mathbf{c}]_\times \frac{1 - \mathbf{a}_1 \cdot \mathbf{a}_2}{||\mathbf{c}||^2} \]  

(14)

where \([\mathbf{c}]_\times\) denotes the skew-symmetric cross product of \(\mathbf{c} = [c_1, c_2, c_3]^T\)

\[
[\mathbf{c}]_\times = \begin{bmatrix}
0 & -c_3 & c_2 \\
c_3 & 0 & -c_1 \\
-c_2 & c_1 & 0
\end{bmatrix}
\]  

(15)

2) Pure rotation

The position vector \(\gamma(t)\) is equal to the previous one \(\gamma(t - 1)\) when a pure rotation acts on, then the translational velocity \(\mathbf{v}(t) = 0\) for all time instances. Accordingly, the translational invariants \(\mathbf{S}_r\), are equal to \([\mathbf{0}, \mathbf{0}, \mathbf{0}]^T\) of length \(N\), where \(\mathbf{0}\) represents the column vector with all zero elements.

3) Pure translation

Under pure translation, the rotational invariants \(\mathbf{S}_r\) will be \([1, 0, 0, 0]^T\), where \(\mathbf{1}\) and \(\mathbf{0}\) are the column vectors of length \(N\) with the entire entries equal to one and zero, respectively.

IV. EXTENSIONS TO MULTIPLE RIGID BODIES

In this section, the mechanism for applying the RRV descriptor on representing motion trajectories of multiple rigid bodies will be introduced.

Firstly, considering \(L\) rigid bodies are free of connection, the integrated descriptors can concatenate the descriptor \(\mathbf{S}_i\) (\(i = 1, \ldots, L\)) for each rigid body as \(\mathbf{S}_{\text{total}} = [\mathbf{S}_1^T, \ldots, \mathbf{S}_L^T]^T\). Beyond this simple case, we focus on more complicated articulated interconnections of multiple rigid bodies in this paper, which can be widely applied in skeleton-based human action recognition. As well known, human action analysis is one of the essential topics in human-robot interaction. With the development of low-cost depth sensors like Kinect, the skeleton of a human body can be extracted from consecutive depth images [21, 22]. The skeleton captured by Kinect consists of the 3D positions of 20 joints \(J_n (n = 1, \ldots, 20)\) as plotted in fig. 4.

Without of loss generality, each connection between two adjacent joints can be assumed as a rigid body. A toy example in fig. 4 indicates the rigid body connecting \(J_1\) and \(J_{11}\), \(\mathbf{b}_i\) is the vector located on the surface of this rigid body and the rotation along \(\mathbf{b}_i\) cannot be derived from 3D positions of the skeleton.

Human actions typically involve different parts of a human body. To effectively integrate our RRV descriptor for multiple rigid bodies in this paper, a hierarchical skeleton could be helpful to produce levels of essential components for representing different actions [23].

A. Hierarchical structure of the skeleton

The whole skeleton of a human body can be decomposed into five components as indicated in fig. 5: Left arm (LA), Right arm (RA), Torso (TS), Left leg (LL) and Right leg (RL). In other words, the action of a human body is the combination of these five hierarchical components. Without loss of generality, we can assume that rigid bodies belong to the same component are correlated to each other while acting a certain action, which means the orientations and positions of these rigid bodies will change simultaneously.

Commonly, the integrated descriptor for each hierarchical component can directly concatenate descriptors of \(L\) rigid bodies together. Fig. 6(a) demonstrates an example of a left arm connected by 4 rigid bodies, then the corresponding concatenated descriptor for representing the motion of this left arm can be written as \(\mathbf{S}_{LA} = [\mathbf{S}_1^T, \mathbf{S}_2^T, \mathbf{S}_3^T, \mathbf{S}_4^T]^T\) and \(\mathbf{S}_{LA} \in \mathbb{R}^{28 \times N}\). However, this simple concatenation will produce ambiguity in some special situations. Considering a man lifts his arm from the left side and front side, the hierarchical descriptors \(\mathbf{S}_{LA}\) for these two actions are the same. Therefore, this confused representation will lead to unreasonable recognition and matching results. Besides, this simply concatenated descriptor also has high dimensions, thus increasing the computational cost during matching and recognition.

![Fig. 4](image1.png)

Fig. 4. An example of the human body’s skeleton extracted by Kinect. The articulated interconnection between two adjacent joints can be assumed as a rigid body.

![Fig. 5](image2.png)

Fig. 5. The hierarchical skeleton of a human body consists of five components: left arm, right arm, torso, left leg and right leg.
B. Virtual rigid body

To improve the descriptive power of the RRV descriptor for characterizing multiple rigid bodies, we will introduce a novel Virtual Rigid Body in this paper. Instead of the reality rigid bodies as shown in fig. 6(a), a virtual rigid body is constructed to provide an alternative representation for each hierarchical component. As indicated in fig. 6(b), we can denote a root joint \( J_{\text{root}} \) and an end joint \( J_{\text{end}} \), respectively. These two joints can be assumed to locate on the surface of this virtual rigid body, and the point \( J_{\text{end}} \) is chosen as the reference point.

Recall that the rigid body is a solid object in which the deformation is not existed. It is important to note that the virtual rigid body proposed here is just an analogical concept, where the distance between \( J_{\text{root}} \) and \( J_{\text{end}} \) will not be limited to a constant value. Let denote the vector connecting these two points as \( \mathbf{b}_1 \). According to the definition of the RRV descriptor, the key step is to estimate 3D rotations of this virtual body and the velocity vectors of \( J_{\text{end}} \). However, the other two principal axes \( \mathbf{b}_2 \) and \( \mathbf{b}_3 \) cannot be defined only given two points on this virtual body. Hence, 3D rotations can be alternatively calculated by \( \mathbf{b}_1 \) across two adjacent time instances as addressed in (14), where rotations along \( \mathbf{b}_2 \) will not be taken into consideration. Given a pair of \( \mathbf{b}_1(t) \) and \( \mathbf{b}_1(t+1) \) at time instant \( t \) and \( t+1 \) as indicated in fig. 6(c), the re-parameterization rotation matrix \( \mathbf{R} \) can be derived. After representing the SRVF of \( J_{\text{end}} \) in the local coordinate system \( \mathbf{R} \), the corresponding RRV descriptor will be determined for this virtual rigid body.

By introducing this concept, the latent spatiotemporal correlations between the root and end joints will be captured, which can offer more motion clues in the description. Similarly, other candidate rigid bodies (black dash lines) in \( LA \) can be established as shown in fig. 6(d), where the root joint is fixed. However, it is important to balance the tradeoff between descriptive power and conciseness of the integrated descriptor. As demonstrated in [24, 25], those acral joints can be denoted as the informative joints, which show more richness in recognizing various actions. Hence, the joint located at the extremity of the skeleton is preferred in constructing the virtual rigid body in each hierarchical component.

Similarly, the descriptors \( S_{R1}, S_{R2}, S_{L1} \) and \( S_{L2} \) for the other four hierarchical components can be built. Then the integrated descriptor for representing the action of the whole skeleton can be expressed as \( S_{\text{human}} = [S_{L1}^t, S_{L2}^t, S_{R1}^t, S_{R2}^t, S_{L1}^t, S_{L2}^t]^T \).

C. Metric between multiple RRV descriptors

Considering the integrated descriptor \( S_{\text{multi}} = [S_1^t, \cdots, S_L^t]^T \) containing \( L \) RRV descriptors, then the distance between two integrated descriptors \( S_{\text{multi}}^p \) and \( S_{\text{multi}}^q \) is the summation of \( d(p,q) \) for each pair \( S'_p \) and \( S'_q \) \((i=1,\cdots,L)\). It can be written as:

\[
d_{\text{multi}}(p,q) = \sum_{k=1}^{L} d_k(p,q) \tag{16}\]

where \( d(p,q) \) has been introduced in (13).

V. RIGID BODY MOTION RECOGNITION

A. Dynamic time warping based recognition

As the rigid body motion trajectories are often parameterized by time series, our RRV descriptor is a time sequence accordingly. To compare the similarity of two time sequences with unequal size, dynamic time warping (DTW) is a popular method. Based on the idea of dynamic programming, the sequences will be nonlinearly warped in a certain constraint, which is to explore the minimum cost path in the distance matrix \( \mathbf{D} \). The constraint in a minimum cost can be written as:

\[
\mathbf{D}(p,q) = d(p,q) + \min\{\mathbf{D}(p-1,q), \mathbf{D}(p,q-1), \mathbf{D}(p-1,q-1)\} \tag{17}\]

where \( d(p,q) \) is the flexible metric as introduced in (13). The minimum accumulative cost of aligning these two sequences will be recorded in \( \mathbf{D}(P,Q) \). For \( L \) multiple rigid bodies, the total cost is

\[
D_{\text{multi}} = \sum_{k=1}^{L} \mathbf{D}(P,Q) \tag{18}\]

As well known, the computation cost of DTW-based recognition method is significantly high as the number of training samples increases. For matching and recognition, a testing sequence will be labeled as the same category as its nearest neighbor (1-NN) in the constraint of minimum aligning cost among all training descriptor sequences.

B. Bag-of-words based recognition

In this paper, a bag-of-words (BoW) based recognition method is proposed to balance the tradeoff between the computational cost and recognition accuracy. It is clear that RRV
descriptor sequences are of unequal size. Recently, BoW [16] approaches become popular since it can provide a feasible technique to match samples with unequal sizes and inter-class variations. According to BoW approach, a single descriptor $S$ of length $m$ can be called a patch. For all given patches from the whole train data, the $K$ cluster centroids can be offline learned by a fast k-means clustering method [26]. These $K$ vectors will be composed of a code book or a dictionary $[W_i]_{i=1\sim K} \in \mathbb{R}^{K\times m}$, where each vector in the dictionary can be denoted as a "visual word". The fundamental idea of BoW is to find the existence of each visual word in a predefined test sample. Hence, given a testing patch, the standard $l_2$-norm from the patch to the $K$ cluster centroid $[W_i]_{i=1\sim K}$ will be calculated, thus to finding the nearest neighbor. Finally, all the patches in a sample will be quantized as a sparse histogram $H \in \mathbb{R}^{K\times1}$, in which the $i$th bin of $H$ represents the occurrence frequency of the word $W_i$. Then the recognition of the original descriptor sequences can be transformed into the classification of normalized histograms. However, temporal information will be neglected during the transformation from time sequences to histogram, which leads to a loss in recognition accuracy.

Recall that the RRV descriptor $S=[S_{\text{r}},S_{\text{v}}]'$ has the rotational invariants $S_{\text{r}} \in \mathbb{R}^{3\times N}$ and the translational invariants $S_{\text{v}} \in \mathbb{R}^{3\times N}$. In this paper, two dictionaries $[W_{\text{r}}^i]_{i=1\sim K_1}$ and $[W_{\text{v}}^i]_{i=1\sim K_2}$ for these two parts will be established separately, where the sizes of two dictionaries are $K_1$ and $K_2$ respectively. As a result, two sparse histograms $H^r \in \mathbb{R}^{K_1\times1}$ and $H^v \in \mathbb{R}^{K_2\times1}$ will be concatenated as $H=[H^r,H^v]'$, and $H \in \mathbb{R}^{(K_1+K_2)\times1}$ is the final representation for the RRV descriptor.

For multiple class classification, we train the nonlinear SVM using the chi-square kernel, which is especially suitable for histogram features. Finally, recognition results will be given based on a one-against-all rule.

VI. EXPERIMENTS

In this section, our RRV descriptor for rigid body motion recognition is performed on different datasets to compare with the previous descriptors and other state-of-the-art approaches.

A. Experimental setup

1) Australian sign language (ASL) dataset

Sign languages are always carried out with two hands, where each palm can be regarded as a rigid body. The 6D motion trajectory of each palm is recorded as shown in fig. 7. The ASL dataset [17] has 95 sign classes, where 9 subjects (3 samples in each subject) for each class are performed by different individuals to improve intra-class variation.

To show the comparative results with the previous ISA descriptor [10, 11] and SoSaLe descriptor [12], we implement them on ASL dataset. The ISA descriptor is invariant to the initial pose, and the initial ISA is chosen as $[0,0,1]'$. Fig. 8 shows an example of the right-hand trajectory of a sign language "all" and the corresponding RRV descriptor. For a specified sign language, the integrated descriptor can be written as $S_{\text{sign}}=[S_{\text{r}}^1,S_{\text{v}}^1]'$, where the subscripts "LH" and "RH" are short for the left hand and right hand respectively.

To demonstrate the robustness and consistency of our approach, the two-fold cross validation of 95 classes are conducted, where 4 subjects are selected as train data and the remaining 5 subjects are chosen for testing. In addition, the proposed metric in (13) and standard $l_2$-norm for measuring the distance between two RRV descriptors are also compared. With regard to the BoW-based recognition, the size of two dictionaries is experimentally set as $K_1=120$ and $K_2=130$.

Fig. 7. The 6D motion trajectory of the right palm, where the reference point is the middle point of the wrist.

Fig. 8. The right hand trajectory of a sign language "all" and the corresponding RRV descriptor. (a) The positions and Euler angles of the right palm recorded in 3D space; (b) the corresponding translational and rotational invariants of the RRV descriptor.
2) MSRAction3D dataset

The MSRAction3D dataset is a popular and publicly available dataset for human action recognition, where many researchers have reported their results on this dataset. This dataset consists of twenty actions: \textit{high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw}. Each action has 10 subjects (2-3 samples in each subject) performed by different individuals. The 3D positions of 20 joints are extracted from the consecutive depth images by using the real-time skeleton tracking algorithm [22]. The frame rate is 15 frames per second, and the samples of the same action have the different size of lengths. The total number of available samples is 557, where the remaining 43 sequences have missing or wrong skeletal data.

It should be noted that our approach only focuses on utilizing the skeletal data to achieve human action recognition rather than the fusion of depth data. To demonstrate the comparative results with the other state-of-the-art skeletal-based works, we first compare our approach with other works using the settings as [18], where 20 actions are grouped into three action subjects AS1, AS2, and AS3 as listed in Table I. The actions in AS1 and AS2 are with similar movements, while those in AS3 are with relative complex movements.

| TABLE I. THREE ACTION SUBJECTS OF MSRActions3D |
|-----------------|-----------------|-----------------|
| AS1             | AS2             | 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      |
| Pick up & throw | Side boxing     | Pick up & throw |

We also follow the more challenging experimental settings as mentioned in [18, 27]. The subjects 1, 3, 5, 7, 9 are chosen for training and subjects 2, 4, 6, 8, 10 are for testing. To illustrate the robustness and consistency of our approach, the cross validation with all possible 5-5 splits (252 of them in total) are conducted.

As suggested in Table II, the following 10 virtual rigid bodies are selected for representing the skeleton of a human body. Specifically, each hierarchical component is represented by two virtual rigid bodies. For BoW-based recognition, the size of two dictionaries is set as $K_1 = 120$ and $K_2 = 180$ respectively.

| TABLE II. ROOT AND END JOINTS OF VIRTUAL RIGID BODIES |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $J_{\text{root}}$ | 3 | 3 | 3 | 3 | 5 | 6 | 7 | 7 | 7 |
| $J_{\text{end}}$   | 11 | 13 | 10 | 12 | 20 | 20 | 19 | 17 | 18 | 16 |

3) Preprocessing

Raw trajectories are usually contaminated with noise and outliers, which motivated us to smooth them before the description. As well is known, the smoothing process may change the shape of input trajectories more or less. To balance the tradeoff between smoothing effects and shape preservation, it is necessary to choose an appropriate smoother. The moving average smoother can provide acceptable performance in alleviating the white noise, whereas it may change trajectory shapes while detecting outliers. Kalman smoother [28] is an available filter which can remain the basic shape of a trajectory as well as smooth the noisy points, and it is applied in this paper. In addition, a number of raw trajectories may record a fraction of stationary sequences at the beginning or end, and these fractions make no contribution in motion characterization but increase the computational cost. Inspired by this, we will remove these stationary data in the preprocessing step.

B. Experimental results

1) Results on ASL dataset

Firstly, the effectiveness of our RRV descriptor will be evaluated on ASL dataset. The recognition results using different descriptors with the DTW-based recognition method and BoW-based recognition method are given. For RRV descriptor, the recognition results provided by the conventional $l_2$-norm is also reported. The distance between two ISA descriptors and two SoSaLe descriptors are also computed by separating the translational invariants and rotational invariants. The mean accuracy and the standard derivation (STD) of all two-fold cross-validation results are recorded in Table III. As shown in the table, the mean and STD result provided by our RRV descriptor significantly outperforms that by ISA descriptor (76.49±2.48%) and SoSaLe descriptor (69.13±2.25%), which mainly attribute to the intra-class similarity and inter-class discriminability of our RRV descriptor. Moreover, compared to conventional $l_2$-norm, the proposed metric will also improve the recognition performance by computing the translational and rotational invariants separately. As demonstrated, the highest recognition accuracy achieves 92.56±2.04% while trajectories are represented by the RRV descriptor and measured by the metric as introduced in (13). Without considering temporal information, the results by BoW-based approach (90.92±2.61%) is slightly inferior to that by the DTW-based method.

| TABLE III. RECOGNITION RESULTS ON ASL DATASET |
|-----------------|-----------------|
| Methods         | Mean±STD(%)     |
| SoSaLe [12] (DTW) | 69.13±2.25     |
| ISA descriptor [10, 11] (DTW) | 76.49±2.48     |
| RRV descriptor ($l_2$-norm+DTW) | 87.45±2.09     |
| RRV descriptor (BoW+SVM) | 90.92±2.61     |
| RRV descriptor (new metric+DTW) | 92.56±2.04     |

Table IV reports the computational cost in terms of the different recognition approaches. By choosing BoW-based recognition approach, the time cost for each testing sample only consumes 24ms. It means the BoW-based approach can balance the tradeoff between the recognition accuracy and the computational cost. In addition, the computational cost provided by RRV descriptor with the DTW-based recognition approach is in the same level as that by previous descriptors.

| TABLE IV. COMPUTATIONAL COST FOR VARIOUS APPROACHES |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Methods         | Mean±STD(%)     |
| SoSaLe [12] (DTW) | 69.13±2.25     |
| ISA descriptor [10, 11] (DTW) | 76.49±2.48     |
| RRV descriptor ($l_2$-norm+DTW) | 87.45±2.09     |
| RRV descriptor (BoW+SVM) | 90.92±2.61     |
| RRV descriptor (new metric+DTW) | 92.56±2.04     |
Recall that the RRV descriptor only calculates first-order time derivatives of discrete motion trajectories, which is robust to noise contamination and outliers than the previous ones. In order to evaluate this, all 6D motion trajectories will be added by white Gaussian noise with signal noise ratio (SNR) from 10 to 50. Figure 9 compares the mean accuracies provided by different descriptors for representing noisy trajectories. As can be seen, the blue triangle and black circle lines are the average values of ISA and SoSaLe descriptors respectively, where both accuracies will drastically decrease as SNR decreases. On the contrary, the results achieved by the RRV descriptor (red square) are more robust in terms of noise contamination.

2) Results on MSRAction3D dataset

Next, we will experimentally evaluate the extensions of the RRV descriptor for representing multiple trajectories on MSRAction3D dataset, which achieves skeleton-based human actions recognition.

Firstly, our approach is compared with other state-of-the-art skeletal-based approaches on three action subjects as listed in Table I. For different subjects AS1, AS2, and AS3, the average recognition accuracies in Table V illustrate that our method (96.4%) outperforms previous approaches. Particularly for the subject AS2 with similar actions, the accuracy achieved by the RRV descriptor with different methods can both achieve the result of 95.5%, which is superior to that by previous approaches.

2.1. Recognition results on three subjects

| Methods         | AS1 (%) | AS2 (%) | AS3 (%) | Average(%) |
|-----------------|---------|---------|---------|------------|
| 3D Bag[18]      | 72.9    | 71.9    | 79.2    | 74.7       |
| HOJ3D[29]       | 88.0    | 85.5    | 63.5    | 79.0       |
| Lie Group[30]   | 95.4    | 83.9    | 98.2    | 92.5       |
| Moving pose[31] | 96.4    | 91.6    | 99.1    | 95.7       |
| RRV(BoW+SVM)    | 93.3    | 95.5    | 97.3    | 95.4       |
| RRV(DTW)        | 95.4    | 95.5    | 98.2    | 96.4       |

Next, we compare the recognition results on the whole MSRAction3D dataset. As provided in Table VI, our approaches still achieve better performance than previous skeleton-based works. The highest recognition accuracy 93.44% is obtained by RRV descriptor, which outperforms the second best (91.70%) by 1.74%. Fig. 10 shows the detailed confusion matrix while the accuracy equals to 93.44%. As observed from the figure, most of the actions can be well classified, whereas the performance of some similar actions (hammer, hand catch, and high throw) is still needed to be improved. Without considering temporal information, the results provided by BoW-based approach (91.21%) is slightly inferior to that by DTW-based method (93.44%). Since some recent researches [32, 33] have reported their recognition accuracy around 94.8% by fusing skeletal features with depth-based features. It should be noted that our paper only focuses on the usage of skeletal features of the human body, thus exploring a novel configuration by representing the skeleton with multiple rigid bodies.

| Methods         | Average (%) |
|-----------------|-------------|
| 3D Bag[18]      | 74.70       |
| HOJ3D[29]       | 75.80       |
| Actionlet[27]   | 82.33       |
| HON4D[34]       | 88.89       |
| Lie Group[30]   | 89.48       |
| Pose Set[35]    | 90.22       |
| Moving pose[31] | 91.70       |
| RRV(BoW+SVM)    | 91.21       |
| RRV(DTW)        | 93.44       |

In addition, to demonstrate the consistency of our approach, Table VII illustrates the cross-validation results of mean accuracies and STD results, where most of the previous works didn’t present this result. As can be seen, our approach achieves a better result of 87.54±3.66%, which outperforms the other proposed approaches.

| Methods         | Mean±STD (%) |
|-----------------|--------------|
| HOJ3D[29]       | 63.5±5.23    |
| HON4D[34]       | 81.88±4.45   |
| RRV(BoW+SVM)    | 84.05±2.83   |
| RRV(DTW)        | 87.54±3.66   |

VII. Conclusion

This paper proposes a novel Rotation and Relative Velocity (RRV) descriptor for representing 6D rigid body motion trajectories by exploring the local translational and rotational invariants. Our descriptor is robust to noise contamination and is invariant to rigid transformation and scaling. Compared to the previous descriptors, the key contribution of our RRV descriptor is that it can capture the spatiotemporal correlations between point trajectory and rigid body’s pose. Moreover, it only involves first-order time derivatives of discrete trajectories, which is simple to construct and insensitive to noise. Instead of the conventional $l_2$-norm, a flexible metric is proposed to measure the distance between two RRV descriptors. In the following, we extend our RRV descriptor to the applications of multiple rigid bodies in the context of skeleton-based human action recognition. The whole skeleton can be firstly
decomposed into five hierarchical components. Another contribution of this paper is proposing a novel virtual rigid body method for alternative representing multiple rigid bodies in each hierarchical. This virtual rigid body can reduce ambiguity and improve discriminability in the description. The effectiveness and consistency of the RRV descriptor is demonstrated by means of rigid body motion recognition studies. The reported experimental results on ASL and MSRAction3D datasets emphasize that the RRV descriptor outperforms the previous ones in terms of recognition accuracy without increasing computational cost.

REFERENCES

[1] C. Cedras and M. Shah, "Motion-based recognition a survey," Image and Vision Computing, vol. 13, pp. 129-155, 1995.
[2] J. P. Ferreira and M. M. Crisóstomo, "Human gait acquisition and characterization," IEEE Transactions On Instrumentation and Measurement, vol. 58, pp. 2979-2988, 2009.
[3] H. Zhou, H. Hu, H. Liu, et al., "Classification of upper limb motion trajectories using shape features," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 42, pp. 970-982, 2012.
[4] S. Wu and Y. F. Li, "On signature invariants for effective motion trajectory recognition," The International Journal of Robotics Research, vol. 27, pp. 895-917, 2008.
[5] I. N. Junejo, E. Dexter, I. Laptev, et al., "View-independent action recognition from temporal self-similarities," IEEE Transactions On Pattern Analysis and Machine Intelligence, vol. 33, pp. 172-185, 2011.
[6] S. Belongie, J. Malik, and J. Puzicha, "Shape context: A new descriptor for shape matching and object recognition," in Proceedings of NIPS, 2000, p. 3.
[7] R. Tello, "Fourier descriptors for computer graphics," IEEE Transactions on Systems, Man and Cybernetics, vol. 25, pp. 861-865, 1995.
[8] Z. Shao and Y. Li, "On Integral Invariants for Effective 3-D Motion Trajectory Matching and Recognition," IEEE transactions on cybernetics, 2015.
[9] J. De Schutter, E. Di Lello, J. F. De Schutter, et al., "Recognition of 6 dof rigid body motion trajectories using a coordinate-free representation," in Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2011, pp. 2071-2078.
[10] J. De Schutter, "Invariant description of rigid body motion trajectories," Journal of Mechanisms and Robotics, vol. 2, p. 011004, 2010.
[11] M. Vochten, T. De Laet, and J. De Schutter, "Comparison of Rigid Body Motion Trajectory Descriptors for Motion Representation and Recognition," in Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2015.
[12] R. Soloperto, M. Saveriano, and D. Lee, "A Bidirectional Invariant Representation of Motion for Gesture Recognition and Reproduction," in Proceedings of International Conference on Robotics and Automation (ICRA), 2015.
[13] M. G. Wagner and B. Ravani, "Curves with rational Frenet-Serret motion," Computer Aided Geometric Design, vol. 15, pp. 79-101, 1997.
[14] A. Srivastava, E. Klassen, S. H. Joshi, et al., "Shape analysis of elastic curves in euclidean spaces," IEEE Transactions On Pattern Analysis and Machine Intelligence, vol. 33, pp. 1415-1428, 2011.
[15] M. Devanne, H. Wannous, S. Berretti, et al., "3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold," IEEE Transactions on Cybernetics, 2014.
[16] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 524-531.
[17] M. W. Kadous, "Temporal classification: Extending the classification paradigm to multivariate time series," The University of New South Wales, 2002.
[18] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 9-14.
[19] J. B. Kuipers, Quaternions and rotation sequences vol. 66: Princeton university, 1999.
[20] D. Q. Huynh, "Metrics for 3D rotations: Comparison and analysis," Journal Of Mathematical Imaging and Vision, vol. 35, pp. 155-164, 2009.
[21] J. Han, L. Shao, D. Xu, et al., "Enhanced computer vision with microsoft kinect sensor: A review," IEEE Transactions on Cybernetics, vol. 43, pp. 1318-1334, 2013.
[22] J. Shotton, T. Sharp, A. Kipman, et al., "Real-time human pose recognition in parts from single depth images," Communications Of The Acm, vol. 56, pp. 116-124, 2013.
[23] N. D. Cornea, D. Silver, and P. Min, "Curve-skeleton properties, applications, and algorithms," *IEEE Transactions On Visualization and Computer Graphics*, pp. 530-548, 2007.

[24] F. Ofli, R. Chaudhry, G. Kurillo, et al., "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *Journal Of Visual Communication and Image Representation*, vol. 25, pp. 24-38, 2014.

[25] G. Chen, D. Clarke, M. Giuliani, et al., "Combining unsupervised learning and discrimination for 3D action recognition," *Signal Processing*, 2014.

[26] J. Z. Lai, T.-J. Huang, and Y.-C. Liaw, "A fast k-means clustering algorithm using cluster center displacement," *Pattern Recognition*, vol. 42, pp. 2551-2556, 2009.

[27] J. Wang, Z. Liu, Y. Wu, et al., "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290-1297.

[28] Y. Feng, M. Ji, J. Xiao, et al., "Mining Spatial-Temporal Patterns and Structural Sparsity for Human Motion Data Denoising," *IEEE transactions on cybernetics*, 2014.

[29] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 20-27.

[30] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 588-595.

[31] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 2752-2759.

[32] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG2 for action recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 465-470.

[33] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 486-491.

[34] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 716-723.

[35] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 915-922.