3D Skeletal Gesture Recognition via Sparse Coding of Time-Warping Invariant Riemannian Trajectories

Xin Liu and Guoying Zhao*

Center for Machine Vision and Signal Analysis, University of Oulu, 90014, Finland.
{xin.liu,guoying.zhao}@oulu.fi

Abstract. 3D skeleton based human representation for gesture recognition has increasingly attracted attention due to its invariance to camera view and environment dynamics. Existing methods typically utilize absolute coordinate to present human motion features. However, gestures are independent of the performer's locations, and the features should be invariant to the body size of performer. Moreover, temporal dynamics can significantly distort the distance metric when comparing and identifying gestures. In this paper, we represent each skeleton as a point in the product space of special orthogonal group $SO3$, which explicitly models the 3D geometric relationships between body parts. Then, a gesture skeletal sequence can be characterized by a trajectory on a Riemannian manifold. Next, we generalize the transported square-root vector field to obtain a re-parametrization invariant metric on the product space of $SO(3)$, therefore, the goal of comparing trajectories in a time-warping invariant manner is realized. Furthermore, we present a sparse coding of skeletal trajectories by explicitly considering the labeling information with each atoms to enforce the discriminant validity of dictionary. Experimental results demonstrate that proposed method has achieved state-of-the-art performance on three challenging benchmarks for gesture recognition.

Keywords: Gesture recognition · Manifold · Sparse coding.

1 Introduction

Human gesture analysis is emerging as a central problem in computer vision applications, such as human-computer interfaces and multimedia information retrieval. 3D skeleton-based modeling is rapidly gaining popularity due to it simplifies the problem caused by replacing monocular RGB camera with more sophisticated sensors such as the Kinect. It can explicitly localize gesture performer and yield the trajectories of human skeleton joints. Compared to RGB data, skeletal data is robust to varied background and is invariant to camera view-point. In the past decade, a considerable number of 3D skeleton-based recognition methods [23, 24, 19, 4, 3, 2, 16, 7, 21, 20, 22, 15, 5, 13, 14] have been proposed. Although there have been significant advancements in this area, accurate

* Corresponding author.

recognition of the human gesture in unconstrained settings still remains challenging. There are two issues need to be thoroughly discussed:

* One important issue in gesture recognition is the feature representation of models to capture variability of 3D human body (skeleton) and its dynamics. Existing methods typically utilize absolute (real world) coordinate to present human motion features. However, activities are independent of performer's locations, and the feature should be invariant to the size of the performer.

* Another issue of human gesture recognition lies in the temporal dynamics. For instance, even the same actions or gestures performed by the same person can have different implementation rates and different starting/ending points, let alone different performers.

A common way to deal with the first problem is to transform all 3D joint coordinates from the world coordinate system to a performer-centric coordinate system by placing the hip center at the origin, but the accuracy heavily depends on the precise positioning of the human hip center. Another solution is to consider the relative geometry between different body parts (bones), such as the Lie Group [19], which utilize rotations and translations (rigid-body transformation) to represent the 3D geometric relationships of body parts. However, the translation is not a scale-invariant representation since the size of skeleton varies from subject to subject.

To account for the second issue, a typical treatment is using the graphical model to describe the presence of sub-states, where the time series are reorganized by a sequential prototype, and the temporal dynamics of gestures are trained as a set of transitions among these prototypes [2]. The typical model is the hidden Markov model (HMM) [22]. However, in these models, the input sequences have to be previously segmented on the basis of specific clustering metrics or discriminative states, which itself is a challenging task. With the development of deep learning, plenty of researches [5, 13, 14] addressing the problem of temporal dynamics by recurrent neural networks (RNN), such as the long short-term memory (LSTM). Although LSTM is a powerful framework for modeling sequential data, it is still arduous to learn the information of the entire sequence with many sub-events. In fact, the most common solution to temporal dynamics is the Dynamic Time Warping (DTW) [19, 7], which needs to choose a nominal temporal alignment, and then all sequences of a category are warped to that alignment. However, the performance of DTW is highly depends on the selection of a reference, which is commonly computed by experience.

Aiming to tackle above issues, in this paper, a novel method for gesture recognition is proposed. The main contributions are summarized as follows:

1) we represent a human skeleton as a point on the product space of special orthogonal group ($SO3$), which is a Riemannian manifold. This representation is independent to the performer's location, and can explicitly models the 3D geometric relationships between body parts using rotations. Then a gesture (skeletal sequences) can be represented by a trajectory composed of these points (see Fig. 1 (d)). The gesture recognition task is formulated as the problem of computing the similarity between the shapes of trajectories.

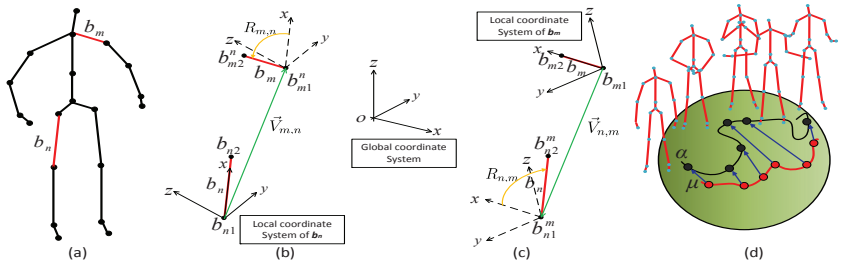

Fig. 1. (a) Illustration of a 3D skeleton, (b) Representation of bone b_m in the local coordinate system of b_n, (c) Representation of b_n in the local coordinate system of b_m, (d) Pictorial of the warped trajectory α on a manifold according to a reference μ.

2) we extend the transported square-root vector field (TSRVF) representation for comparing trajectories on the product space of $SO(3) \times \cdots \times SO(3)$. Therefore, the temporal dynamic issue of gesture recognition can be solved by this time-warping invariant feature.

3) we present a sparse coding of skeletal trajectories by explicitly considering the labeling information with each atom to enforce the discriminant validity of dictionary. The comparison experimental results on three challenging datasets demonstrated the proposed method have achieved state-of-the-art performances.

2 Related works

Over the last few years, plenty of 3D skeletal human gesture recognition models have been explored in various routines. In this section, we limited our review on the relevant manifold-based solutions. A representative work is the Lie group [19], which utilized the special Euclidean (Lie) group $SE(3)$ to characterize the 3D geometric relationships among body parts. A convenient way of analyzing Lie group is to embed them into Euclidean spaces, with the embedding typically obtained by flattening the manifold via tangent spaces, such as the Lie algebra $\mathfrak{se}(3)$ at the tangent space identity I_4. In that way, former classification tasks in manifold curve space are converted into the classification problems in a typical vector space. Then, the authors of [19] employed the DTW and Fourier temporal pyramid (FTP) to deal with the temporal dynamics issues of gesture recognition. However, as discussed in Section 1, the success of DTW is heavily related to the choice of the nominal temporal alignment empirically. And the FTP is restricted by the width of the time window and can only utilize limited contextual information [5]. Following the same representation, Anirudh *et al.* [3] introduced the framework of transported square-root velocity fields (TSRVF) [18] to encode trajectories lying on Lie groups, as such, the distance between two trajectories is invariant to identical time warping. Since the final feature is a high-dimensional vector, the principal component analysis (PCA) is used to reduce the dimension and learn the basis (dictionary) for representation. While

PCA is an unsupervised model and thus the discriminant of dictionary cannot be boosted through a labeled training. Based on the square root velocity (SRV) framework [17], in [4], trajectories are transported to a reference tangent space attached to the Kendall's shape space at a fixed point, which may introduce distortions in the case points are not close to the reference point. In [8], Ho *et al.* proposed a general framework for sparse coding and dictionary learning on Riemannian manifolds. Different to [17] which using the fixed point for embedding, the [8] working on the tangent bundle, namely, each point of manifold is coded on its attached tangent space into which the atoms are mapped.

3 Product Space of $SO(3)$ for 3D Skeleton Representation

Inspired by the rigid body kinematics, any rigid body displacement can be realized by a rotation about an axis combined with a translation parallel to that axis. This 3D rigid body displacements forms a Lie group, which is generally referred to as $SE(3)$, the special Euclidean group in three dimensions:

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

where $R \in SO(3)$ is a point in the special orthogonal group $SO(3)$, denotes the rotation matrix, and $\boldsymbol{v} \in \mathbb{R}^3$ denotes the translation vector.

The human skeleton can be modeled by an articulated system of rigid segments connected by joints. As such, the relative geometry between a pair of body parts (bones) can be represented as a point in $SE(3)$. More specifically, given a pair of bones b_m and b_n, their relative geometry can be represented in a local coordinate system attached to other [19]. Let $b_{i1} \in \mathbb{R}^3$, $b_{i2} \in \mathbb{R}^3$ denote the starting and ending points of bones b_i respectively. The local coordinate system of bone b_n is calculated by rotating with minimum rotation and translating the global coordinate system so that b_{n1} act as the origin and b_n coincides with the $x-$axis, Fig. 1 give an example to explain this pictorially. As such, at time t, the representation of bone b_m in the local coordinate system of b_n (Fig. 1 (b)), the starting point $b^n_{m1}(t) \in \mathbb{R}^3$ and ending point $b^n_{m2}(t) \in \mathbb{R}^3$ are given by

$$\begin{bmatrix} b^n_{m1}(t) & b^n_{m2}(t) \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} R_{m,n}(t) & \boldsymbol{v}_{m,n}(t) \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & l_m \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \tag{2}$$

where $R_{m,n}(t)$ and $\boldsymbol{v}_{m,n}(t)$ respectively denote the rotation and translation measured in the local coordinate system attached to b_n, and l_m is the length of b_m. According to the theory of rigid body kinematics, the lengths of bones do not vary with time, thus, the relative geometry of b_m and b_n can be described by

$$P_{m,n}(t) = \begin{bmatrix} R_{m,n}(t) & \boldsymbol{v}_{m,n}(t) \\ 1 & 1 \end{bmatrix} \in SE(3), \quad P_{n,m}(t) = \begin{bmatrix} R_{n,m}(t) & \boldsymbol{v}_{n,m}(t) \\ 1 & 1 \end{bmatrix} \in SE(3) \tag{3}$$

One restriction of this motion feature is the translation $\boldsymbol{v}$ is relative to the size of performer (subject). But as we known it is very important to obtain a scale-invariant skeletal representation for recognition task in an unconstrained environment. To remove the skeletons scaling variability, in this paper, we discard the translation from motion representation, then the relative geometry of b_m and b_n at time t can be described by rotations $R_{m,n}(t)$ and $R_{n,m}(t)$, and expressed as elements of $SO(3)$. Then, let M denotes the number of bones, the resulting feature for an entire human skeleton is interpreted by the relative geometry between all pairs of bones, as a point $C(t) = (R_{1,2}(t), R_{2,1}(t), \ldots, R_{M-1,M}(t), R_{M,M-1}(t))$ on the curved product space (see Fig. 1 (d)) of $SO(3) \times \cdots \times SO(3)$, and the number of $SO(3)$ is $2C_M^2$, where C_M^2 is the combination formula.

4 Trajectories Identification on Riemannian Manifold

As presented above, gesture recognition is formulated as the problem of computing the similarity between shapes of trajectories. The basis for these comparability determinations are related to a distance function on the shape space.

To be specific, let α denote a smooth oriented curve (trajectory) on a Riemannian manifold M, and let $\mathcal{M}$ denote the set of all such trajectories: $\mathcal{M} = \{\alpha : [0,1] \to M | \alpha \text{ is smooth}\}$. Re-parameterizations will be represented by increasing diffeomorphisms $\gamma : [0,1] \to [0,1]$, and the set of all these orientation preserving diffeomorphisms is denoted by $\Gamma = \{\gamma \to [0,1]\}$. In fact, γ plays the role of a time-warping function, where $\gamma(0) = 0, \gamma(1) = 1$ so that preserve the end points of the curve. More specifically, if α in the form of time observations $\alpha(t_1), ..., \alpha(t_n)$, is a trajectory on M, the composition $\alpha \circ \gamma$ in the form of time-warped trajectory $\alpha(\gamma(t_1)), ..., \alpha(\gamma(t_n))$, is also a trajectory that goes through the same sequences of points as α but at the evolution rate governed by γ [18].

For classify trajectories, a metric is needed to describe the variability of a class of trajectories and to quantify the information contained within a trajectory. A directly and commonly solution is to calculate point-wise difference, since M is a Riemannian manifold, we have a natural distance d_m between points on M [18]. Then, the distance d_x between any two trajectories: $\alpha_1, \alpha_2 : [0,1] \to M$:

$$d_x(\alpha_1, \alpha_2) = \int_0^1 d_m\left(\alpha_1(t), \alpha_2(t)\right) dt \tag{4}$$

Although this quantity describes a natural extension of d_m from M to $M^{[0,1]}$, it suffers from the issue that $d_x(\alpha_1, \alpha_2) \neq d_x(\alpha_1 \circ \gamma_1, \alpha_2 \circ \gamma_2)$. As discussed in the Section 1, in the task of recognition, the temporal dynamics is a key issue that need to be solved when a trajectory (gesture) α is observed as $\alpha \circ \gamma$, for a random temporal evolution γ. That is, for arbitrary temporal re-parametrizations γ_1, γ_2 and arbitrary trajectories α_1, α_2, a distance $d(\cdot, \cdot)$ is wanted that enable

$$d(\alpha_1, \alpha_2) = d(\alpha_1 \circ \gamma_1, \alpha_2 \circ \gamma_2) \tag{5}$$

A distance that is particularly well-suited for our goal is the one used in the Square Root Velocity (SRV) framework [17]. Based on the concept of elastic

trajectories in [17], Su [18] proposed a Transported Square-Root Vector Field (TSRVF) to represent trajectories, and the original Euclidean metric based SRV has been generalized to the manifold space based framework. Specifically, for a smooth trajectory $\alpha \in \mathcal{M}$, the TSRVF is a parallel transport of a scaled velocity vector field of α to a reference point $c \in M$ according to

$$h_\alpha(t) = \frac{\dot{\alpha}(t)_{\alpha(t)\to c}}{\sqrt{|\dot{\alpha}(t)|}} \in T_c(M) \tag{6}$$

where $\dot{\alpha}(t)$ is the velocity vector along the trajectory at time t, and $\dot{\alpha}(t)_{\alpha(t)\to c}$ is its transport from the point $\alpha(t)$ to c along a geodesic path, and $|\cdot|$ denotes the norm related to the Riemannian metric on M and $T_c(M)$ denotes the tangent space of M at c. Especially, when $|\dot{\alpha}(t)| = 0$, $h_\alpha(t) = 0 \in T_c(M)$. Since α is smooth, so is the vector field h_α. Let $\mathcal{H} \subset T_c(M)^{[0,1]}$ be the set of smooth curves in $T_c(M)$ obtained as TSRVFs of trajectories in M, $\mathcal{H} = \{h_\alpha | \alpha \in \mathcal{M}\}$ [18]. By means of TSRVF, two trajectories such as α_1 and α_2, can be mapped into the tangent space $T_c(M)$, as two corresponding TSRVFs, h_{α_1} and h_{α_2}. The distance among them can be measured by ℓ_2-norm on the typical vector space

$$d_h(h_{\alpha_1}, h_{\alpha_2}) = \sqrt{\int_0^1 |h_{\alpha_1}(t) - h_{\alpha_2}(t)|^2 dt} \tag{7}$$

The motivation of TSRVF representation comes from the following fact. If a trajectory α is warped by γ, to result in $\alpha \circ \gamma$, the TSRVF of $\alpha \circ \gamma$ is given by

$$h_{\alpha\circ\gamma}(t) = h_\alpha(\gamma(t))\sqrt{\dot{\gamma}(t)} \tag{8}$$

Then, for any $\alpha_1, \alpha_2 \in \mathcal{M}$ and $\gamma \in \Gamma$, the distance d_h satisfies

$$d_h(h_{\alpha_1\circ\gamma}, h_{\alpha_2\circ\gamma}) = \sqrt{\int_0^1 |h_{\alpha_1}(s) - h_{\alpha_2}(s)|^2 ds} = d_h(h_{\alpha_1}, h_{\alpha_2}) \tag{9}$$

where $s = \gamma(t)$. For the proof of equality, we refer the interested reader to [18, 17]. From the geometric point of view, this equality implies that the action of Γ on $\mathcal{H}$ under the ℓ_2 metric is by isometries. It enable us to develop a fully invariant distance to time-warping and use it to properly register trajectories [18]. Also, this invariability in execution rates is crucial for statistical analyses, such as sample means and covariance. Then, we define the equivalence class $[h_\alpha]$ (or the notation $[\alpha]$) to denote the set of all trajectories that are equivalent to a given $h_\alpha \in \mathcal{H}$ (or $\alpha \in \mathcal{M}$).

$$[h_\alpha] = \{h_{\alpha\circ\gamma} | \gamma \in \Gamma\} \tag{10}$$

Clearly, such an equivalent class $[h_\alpha]$ (or $[\alpha]$) is associated with a category of gesture. In this framework, the task of comparison two trajectories is performed by comparing their equivalence classes, in other words, an optimal re-parametrization γ^* is need to be found to minimize the cost function

$d_h(h_{\alpha_1}, h_{\alpha_2 \circ \gamma})$. Let $\mathcal{H}/\sim$ be the corresponding quotient space, this can be bijectively identified with the set $\mathcal{M}/\sim$ using $[h_\alpha] \mapsto [\alpha]$ [3]. The distance d_s on $\mathcal{H}/\sim$ (or $\mathcal{M}/\sim$) is the shortest d_h distance between equivalence classes in $\mathcal{H}$ [18], given by:

$$\begin{aligned} d_s([\alpha_1],[\alpha_2]) \equiv d_s([h_{\alpha_1}],[h_{\alpha_2}]) &= \inf_{\gamma \in \Gamma} d_h(h_{\alpha_1}, h_{\alpha_2 \circ \gamma}) \\ = \inf_{\gamma \in \Gamma} \left(\int_0^1 |h_{\alpha_1}(t) - h_{\alpha_2}(\gamma(t)) \sqrt{\dot{\gamma}(t)}|^2 dt \right)^{1/2} & \end{aligned} \tag{11}$$

In practice, the minimization over Γ is solved for using dynamic programming [18, 17]. One important parameter of TSRVF is the reference point c, which should remain unchanged in the whole process of computing. Since the selection of c can potentially affect the results, typically, a point is a natural candidate for c if most of trajectories pass close to that one. In this paper, the Karcher mean [11] as Riemannian center of mass is selected, since it is equally distant from all the points thereby minimizing the possible distortions. Given a set $\{\alpha_i(t)_{t=1,..,n}\}_{i=1}^m$ of sequences (trajectories), its Karcher mean $\mu(t)$ is calculated using the TSRVF representation with respect to d_s in $\mathcal{H}/\sim$, defined as

$$h_\mu = \arg \min_{[h_\alpha] \in \mathcal{H}/\sim} \sum_{i=1}^m d_s([h_\alpha],[h_{\alpha_i}])^2 \tag{12}$$

As a result, each trajectory is recursively aligned to the mean $\mu(t)$, thus, another output of Karcher mean computing is the set of aligned trajectories $\{\tilde{\alpha}_i(t)_{t=1,...,n}\}_{i=1}^m$. For each aligned trajectory $\tilde{\alpha}_i(t)$ at time t, the shooting vector $v_i(t) \in T_{\mu(t)}(M)$ is computed so that a geodesic that goes from $\mu(t)$ to $\tilde{\alpha}_i(t)$ in unit time [18] with the initial velocity $v_i(t)$

$$v_i(t) = \exp^{-1}_{\mu(t)}(\tilde{\alpha}_i(t)) \tag{13}$$

Then, the combined shooting vectors $V(i) = [v_i(1)^{\mathrm{T}}\ v_i(2)^{\mathrm{T}}\ ...\ v_i(n)^{\mathrm{T}}]^{\mathrm{T}}$ is the final feature of a trajectory α_i.

5 Discriminative Sparse Coding of Riemannian Trajectories

Since the final feature of a trajectory (gesture sequence) lies on a high dimensional vector, a common solution is to utilize the principal component analysis (PCA) to reduce the dimension and learn the basis for representation, such as [18, 17] did. As we know, PCA is an unsupervised learning model without labeled training. Compared to the component analysis techniques, the sparse coding model with labeled training has superior capability to capture inherent relationship among the input data and label information. To the best of our knowledge, few manifold representation-based models considered the connection between the labels and the dictionary learning. In this paper, we try to associate

label information with each dictionary atom to enforce the discriminability of sparse codes during the dictionary learning process.

Given a set of observations (feature vectors of gestures) $\mathcal{Y} = \{y_i\}_{i=1}^N$, where $y_i \in \mathbb{R}^n$, and let $\mathcal{D} = \{d_i\}_{i=1}^K$ be a set of vectors in $\mathbb{R}^n$ denoting a dictionary of K atoms. The learning of dictionary $\mathcal{D}$ for sparse representation of $\mathcal{Y}$ can be expressed as

$$< \mathcal{D}, X >= \arg\min_{\mathcal{D},X} \|\mathcal{Y} - \mathcal{D}X\|_2^2 \qquad s.t.\ \forall i,\ \|x_i\|_0 \leq T \tag{14}$$

where $X = [x_1, ..., x_N] \in \mathbb{R}^{K\times N}$ represents the sparse codes of observation $\mathcal{Y}$, and T is a sparsity constraint factor. The construction of D is achieved by minimizing the reconstruction error $\|\mathcal{Y} - \mathcal{D}X\|_2^2$, and satisfying the sparsity constraints. The K-SVD [1] algorithm is a commonly used solution to (14).

Inspired by [25, 10], the classification error and label consistency regularization are introduced into the objective function:

$$\begin{aligned} < \mathcal{D}, W, A, X >= \arg\min_{\mathcal{D},W,A,X} & \|\mathcal{Y} - \mathcal{D}X\|_2^2 + \beta\|L - WX\|_2^2 \\ & + \tau\|Q - AX\|_2^2 \qquad s.t.\ \forall i,\ \|x_i\|_0 \leq T \end{aligned} \tag{15}$$

where $W \in \mathbb{R}^{C\times K}$ denotes the classifier parameters, and C is the number of categories. $L = [l_1, ..., l_N] \in \mathbb{R}^{C\times N}$ represents the class labels of observation $\mathcal{Y}$, and $l_i = [0, ..., 1, ..., 0]^{\mathrm{T}} \in \mathbb{R}^C$ is a label vector corresponding to an observation y_i, where the nonzero position (index) indicates the class of y_i. Then, the additional term $\|L - WX\|_2^2$ denotes the classification error for label information.

For the last term $\|Q - AX\|_2^2$, where $Q = [q_1, ..., q_N] \in \mathbb{R}^{K\times N}$ and $q_i = [0, ..., 1, ..., 1, ..., 0]^{\mathrm{T}} \in \mathbb{R}^K$ is a sparse code corresponding to an observation y_i for classification, the purpose of setting nonzero elements is to enforce the "discriminative" of sparse codes [10]. Specifically, the nonzero elements of q_i occur at those indices where the corresponding dictionary atom d_n share the same label with the observation y_i. The A denotes a $K \times K$ transformation matrix, which is utilized to transform the original sparse code x to be a discriminative one. Thus, the term $\|Q - AX\|_2^2$ represents the discriminative sparse code error, which enforces that the transformed sparse codes AX approximate the discriminative sparse codes Q. It forces the signals from the same class to have similar sparse representations. β and τ are regularization parameters which control the relative contributions of the corresponding terms. Equation (15) can be rewritten as:

$$< \mathcal{D}, W, A, X >= \arg\min_{\mathcal{D},W,A,X} \left\| \begin{pmatrix} \mathcal{Y} \\ \sqrt{\beta}L \\ \sqrt{\tau}Q \end{pmatrix} - \begin{pmatrix} \mathcal{D} \\ \sqrt{\beta}W \\ \sqrt{\tau}A \end{pmatrix} X \right\|_2^2 \quad s.t.\ \forall i,\ \|x_i\|_0 \leq T \tag{16}$$

Let $\mathcal{Y}' = (\mathcal{Y}^{\mathrm{T}}, \sqrt{\beta}L^{\mathrm{T}}, \sqrt{\tau}Q^{\mathrm{T}})^{\mathrm{T}}$, $\mathcal{D}' = (\mathcal{Y}^{\mathrm{T}}, \sqrt{\beta}W^{\mathrm{T}}, \sqrt{\tau}A^{\mathrm{T}})^{\mathrm{T}}$. Then, the optimization of Equation (16) is equivalent to solving the (14) (replace $\mathcal{Y}$ and $\mathcal{D}$ with $\mathcal{Y}'$ and $\mathcal{D}'$ respectively), this is just the problem that K-SVD [1] solves. In this paper, a similar initialization and optimization solution of K-SVD described in [10] is adopted. For parameter settings, the maximal iteration equals to 60, the sparsity factor $T = 50$ is used, and β and τ are set to 1.0 in our experiments.

6 Experiments

In this section, the proposed 3D skeletal gesture recognition method is evaluated in comparison to state-of-the-art methods using three public datasets, namely, ChaLearn 2014 gesture [6], MSR Action3D [12], and UTKinect-Action3D [23].

In order to testify the effectiveness of the proposed method, eighteen state-of-the-arts are compared, we simply divided these methods into three groups. The first group is the methods most related to us, including four Lie group based algorithms, the Lie group using DTW [19] (Lie group-DTW), Lie group with TSRVF [18] (Lie group-TSRVF) and using PCA for dimensionality reduction [3] (Lie group-TSRVF-PCA), and K-SVD for sparse coding [1] (Lie group-TSRVF-KSVD). And also including two TSRVF related methods, the body part features with SRV and k-nearest neighbors clustering[4] (SRV-KNN), TSRVF on Kendall's shape space [2] (Kendall-TSRVF). The methods in second group are based on classic feature representations, like histogram of 3D joints (HOJ3D) [23], Eigen-Joints [24], actionlet ensemble (Actionlet) [20], histogram of oriented 4D normals (HON4D) [16], rotation and relative velocity with DTW (RVV+DTW) [7], naive Bayes nearest neighbor (NBNN) [21]. The last group including seven deep learning methods, namely the convolutional neural network based ModDrop (CNN) [15], HMM with deep belief network (HMM-DBN) [22], LSTM [9], hierarchical recurrent neural network (HBRNN) [5], spatio-temporal LSTM with trust gates (ST-LSTM-TG) [13], and global context-aware attention LSTM (GCA-LSTM) [14]. The baseline results are reported from their original papers.

To verify the effectiveness of the TSRVF on product space of $SO(3) \times \cdots \times SO(3)$ (SO3-TSRVF), we present its discriminative performance without any further step (such as PCA or sparse coding) on three datasets. For comparison of dictionary learning ability, we also report the results of the classic coding such as K-SVD [1] (SO3-TSRVF-KSVD) and the proposed sparse coding scheme (SO3-TSRVF-SC). In order to fairly comparison, we follow the same classification setup as in [19, 3, 2, 18, 1] , namely, we utilized an one-vs-all linear SVM classifier (the parameter C set to 1.0). All experiments are carried out on an Intel Xeon CPU E5-2650 PC with a NVIDIA Tesla K80 GPU.

The ChaLearn 2014 [6] is a gesture dataset with multi-modality data, including audio, RGB, depth, human body mask maps, and 3D skeletal joints. This dataset collects 13585 gesture video segments (Italian cultural gesture) from 20 classes. We follow the evaluation protocol provided by the dataset which assigns 7754 gesture sequences for training, 3362 sequences for validation, and 2742 sequences for testing. The detailed comparison with other approaches is shown in Table 1 (second column). It can be seen that the proposed method achieves the highest recognition accuracy as 93.2%. Compared to Lie group based methods, the effectiveness of SO3-TSRVF has been proved by the experimental results. It is noted that Lie group-DTW [19] is only 79.2%, this is due to the performance of DTW is highly depends on the reference sequences for each category, and that empiric selection task turn to difficult as the size of dataset get larger. It also can be observed that the accuracy of the LSTM [9] is 6 percents less than the proposed method. Although LSTM is designed for perceiving the contextual

Table 1. Comparison Of Recognition Accuracy (%) With Existing 3D Skeleton-Based Methods on ChaLearn 2014 [6], MSR Action3D [12] and UTKinect-Action3D [23] Datasets (best: bold, second best: underline).

Methods	ChaLearn 2014	MSR Action3D	UTKinect-Action3D
Lie group-DTW [19]	79.2	92.5	97.1
Lie group-TSRVF [18]	91.8	87.7	94.5
Lie group-TSRVF-PCA [3]	90.4	88.3	94.9
Lie group-TSRVF-KSVD [1]	91.5	87.6	92.7
SRV-KNN [4]	-	92.1	91.5
Kendall-TSRVF [2]	-	89.9	89.8
EigenJoints [24]	59.3	82.3	92.4
Actionlet [20]*	-	88.2	-
HOJ3D [23]	-	78.9	90.9
HON4D [16]*	-	88.9	90.9
RVV-DTW [7]	-	93.4	-
NBNN [21]	-	**94.8**	98.0
ModDrop (CNN) [15]*	<u>93.1</u>	-	-
HMM-DBN [22]	83.6	82.0	-
LSTM [9]	87.1	88.9	72.7
HBRNN [5]	-	94.5	-
ST-LSTM-TG [13]	92.0	**94.8**	97.0
GCA-LSTM [14]	-	-	**98.5**
Ours (SO3-TSRVF)	92.1	93.4	96.8
Ours (SO3-TSRVF-KSVD)	92.8	93.7	97.2
Ours (SO3-TSRVF-SC)	**93.2**	<u>94.6</u>	<u>98.1</u>

* The method use skeleton and RGB-D data.

information, it is still challenging to model the sequence with temporal dynamics, especially when training data is limited. It is noted that the ModDrop [15] ranked the first place in Looking at People challenge [6]. While our method can achieve a higher score than ModDrop but without using RGB-D and audio data.

The MSR Action3D [12] is a commonly used dataset, where actions are highly similar to each other and have typical large temporal misalignments. This dataset comprises of 567 pre-segmented action instances, and 10 people performing 20 classes of actions. For a fair comparison, the same evaluation protocol, namely the cross-subject testing as described in [12] is followed, where half of the subjects are used for training (subjects number 1, 3, 5, 7, 9) and the remainder for testing (2, 4, 6, 8, 10). We compare the proposed method with the state-of-the-arts, the recognition accuracies on MSR Action3D dataset are recorded in Table 1 (third column). We can see that the proposed method achieves better performance than Lie group based and classical feature representation approaches. And again, the performance of proposed sparse coding is superior than K-SVD and PCA based coding methods. Actually, the recognition accuracy of the proposed is only 0.2% inferior to the NBNN [21] and ST-LSTM-TG [13], which are recently proposed.

The UTKinect-Action3D [23] is a difficult benchmark due to its high intra-class variations. This dataset collects 10 types of actions using the Kinect. We follow [23] and use the *Leave-One-Sequence-Out Cross Validation* setting which selects each sequence as the testing sample in turn, regards others as training samples and calculates the average (20 rounds of testing) recognition rate. Table 1 (fourth column) reports the comparisons of the proposed to state-of-the-art methods. Obviously, our approach outperforms other methods except the GCA-LSTM [14] which is a sophisticated deep learning model proposed recently.

7 Conclusion

In this paper, a new human gesture recognition method is proposed. We represented a 3D human skeleton as a point in the product space of special orthogonal group $SO3$, as such, a human gesture can be characterized as a trajectory in the Riemannian manifold space. To consider re-parametrization invariance properties for trajectory analysis, we generalize the transported square-root vector field to obtain a time-warping invariant metric for comparing trajectories. Moreover, a sparse coding scheme of skeletal trajectories is proposed by thoroughly considering the labeling information with each atom to enforce the discriminant validity of dictionary. Experiments demonstrate that proposed method has achieved state-of-the-art performances. Possible directions for future work include studying on an end-to-end deep network architecture in the manifold space to handle the issues of 3D skeletal gesture recognition.

Acknowledgments. This work is supported by Academy of Finland, Tekes Fidipro Program, Infotech, Tekniikan Edistamissaatio, and Nokia Foundation.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
2. Amor, B.B., Su, J., Srivastava, A.: Action recognition using rate-invariant analysis of skeletal shape trajectories. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 1–13 (2016)
3. Anirudh, R., Turaga, P., Su, J., Srivastava, A.: Elastic functional coding of human actions: From vector-fields to latent variables. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 3147–3155 (2015)
4. Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., Del Bimbo, A.: 3D human action recognition by shape analysis of motion trajectories on Riemannian manifold. IEEE Trans. Cybern. **45**(7), 1340–1352 (2015)
5. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 1110–1118. IEEE (2015)
6. Escalera, S., Baró, X., Gonzalez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. In: Proc. Eur. Conf. Comput. Vis. Workshops. pp. 459–473. Springer (2014)

7. Guo, Y., Li, Y., Shao, Z.: RRV: A spatiotemporal descriptor for rigid body motion recognition. IEEE Trans. Cybern. (2017)
8. Ho, J., Xie, Y., Vemuri, B.: On a nonlinear generalization of sparse coding and dictionary learning. In: Proc. Int. Conf. Mach. Learn. pp. 1480–1488 (2013)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Jiang, Z., Lin, Z., Davis, L.S.: Label consistent K-SVD: Learning a discriminative dictionary for recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2651–2664 (2013)
11. Karcher, H.: Riemannian center of mass and mollifier smoothing. Commun. Pure Appl. Math. **30**(5), 509–541 (1977)
12. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops. pp. 9–14. IEEE (2010)
13. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Proc. Eur. Conf. Comput. Vis. pp. 816–833. Springer (2016)
14. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention LSTM networks for 3D action recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 1647–1656 (2017)
15. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: ModDrop: Adaptive multi-modal gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8), 1692–1706 (2016)
16. Oreifej, O., Liu, Z.: HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 716–723 (2013)
17. Srivastava, A., Klassen, E., Joshi, S.H., Jermyn, I.H.: Shape analysis of elastic curves in Euclidean spaces. IEEE Trans. Pattern Anal. Mach. Intell. **33**(7), 1415–1428 (2011)
18. Su, J., Kurtek, S., Klassen, E., Srivastava, A.: Statistical analysis of trajectories on Riemannian manifolds: bird migration, hurricane tracking and video surveillance. Ann. Appl. Stat. pp. 530–552 (2014)
19. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3D skeletons as points in a Lie group. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 588–595. IEEE (2014)
20. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3D human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 914–927 (2014)
21. Weng, J., Weng, C., Yuan, J.: Spatio-temporal naive-bayes nearest-neighbor (ST-NBNN) for skeleton-based action recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2017)
22. Wu, D., Shao, L.: Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 724–731. IEEE (2014)
23. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops. pp. 20–27. IEEE (2012)
24. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops. pp. 14–19. IEEE (2012)
25. Zhang, Q., Li, B.: Discriminative K-SVD for dictionary learning in face recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 2691–2698. IEEE (2010)