This CVPR2015 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Elastic Functional Coding of Human Actions: From Vector-Fields to Latent Variables

Rushil Anirudh[1,2], Pavan Turaga[2,1], Jingyong Su[3], and Anuj Srivastava[4]

[1]School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ
[2]School of Arts, Media and Engineering, Arizona State University, Tempe, AZ
[3]Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX
[4]Department of Statistics, Florida State University, Tallahassee, FL

{ranirudh@asu.edu, pturaga@asu.edu, jingyong.su@ttu.edu, anuj@stat.fsu.edu}

Abstract

Human activities observed from visual sensors often give rise to a sequence of smoothly varying features. In many cases, the space of features can be formally defined as a manifold, where the action becomes a trajectory on the manifold. Such trajectories are high dimensional in addition to being non-linear, which can severely limit computations on them. We also argue that by their nature, human actions themselves lie on a much lower dimensional manifold compared to the high dimensional feature space. Learning an accurate low dimensional embedding for actions could have a huge impact in the areas of efficient search and retrieval, visualization, learning, and recognition. Traditional manifold learning addresses this problem for static points in $\mathbb{R}^n$, but its extension to trajectories on Riemannian manifolds is non-trivial and has remained unexplored. The challenge arises due to the inherent non-linearity, and temporal variability that can significantly distort the distance metric between trajectories. To address these issues we use the transport square-root velocity function (TSRVF) space, a recently proposed representation that provides a metric which has favorable theoretical properties such as invariance to group action. We propose to learn the low dimensional embedding with a manifold functional variant of principal component analysis (mfPCA). We show that mfPCA effectively models the manifold trajectories in several applications such as action recognition, clustering and diverse sequence sampling while reducing the dimensionality by a factor of $\sim 250\times$. The mfPCA features can also be reconstructed back to the original manifold to allow for easy visualization of the latent variable space.

Rushil Anirudh & Pavan Turaga were supported by NSF CCF CIF grant #1320267. Anuj Srivastava was supported by NSF CCF CIF grant #1319658.

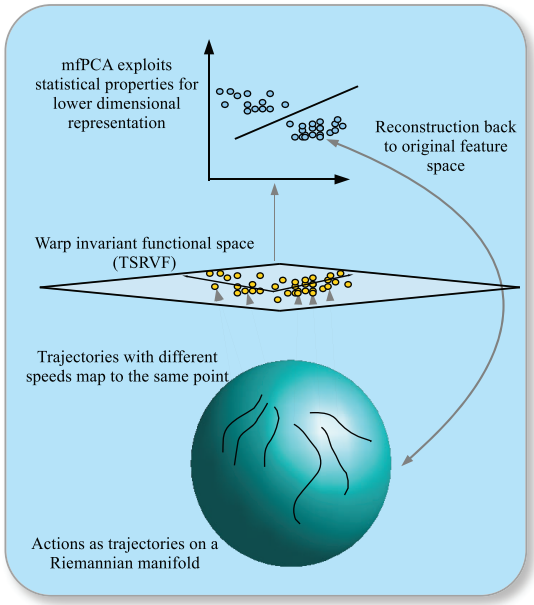


Figure 1: Overview of our work

1. Introduction

There have been significant advances in understanding differential geometric properties of image and video features in vision and robotics. Examples include activity recognition [25, 4, 29], medical image analysis [6], and shape analysis [21]. Some of the popular non-Euclidean features used for activity analysis include shape silhouettes on the Kendall's shape space [28], pairwise transformations of skeletal joints on $SE(3)\times SE(3)\cdots\times SE(3)$ [29], linear dynamical system on the Grassmann manifold [25], and histogram of oriented optical flow (HOOF) on a hyper-sphere [4]. A commonly occurring theme in many applications is

the need to *represent, compare, and manipulate* such representations in a manner that respects certain constraints.

One such constraint is the need for invariance with regard to temporal re-parameterization (or warping) which can distort distance measures significantly, especially in the context of human activities. The most common way to solve for the mis-alignment problem is to use dynamic time warping (DTW) which originally found its use in speech processing [2]. However, DTW behaves as a similarity measure instead of a true distance metric in that it does not naturally allow the estimation of statistical measures such as mean and variance of action trajectories. We seek a representation that is highly discriminative of different classes while factoring out temporal warping to reduce the variability within classes. Learning such a representation is complicated when the features extracted are non-Euclidean (i.e. they do not obey conventional properties of the Euclidean space). Finally, typical representations for action recognition tend to be extremely high dimensional in part because the features are extracted per-frame and stacked. Any computation on such non-linear trajectories is expensive.

By the nature of human movements, actions don't span the entire high dimensional feature space, therefore it is reasonable to assume that the actions themselves lie on a much lower dimensional manifold. A lower dimensional embedding will provide a robust, computationally efficient, and intuitive framework for analysis of actions. In this paper, we address these issues by studying the statistical properties of trajectories on Riemannian manifolds. We consider two types of features for human actions which lie on different manifolds - shape silhouettes on the Grassmann manifold [25] and skeletal joints as points on the product space $SE(3) \times \cdots \times SE(3)$[29]. The techniques proposed in this work can also be applied to features such as relative joint positions of skeletal data that lie in $\mathbb{R}^n$. We show that the lower dimensional embedding can accurately recognize actions on benchmark datasets such as UTKinect[31], Florence3D [18] and MSR Actions3D [14] better than the original features, on a significantly lower dimensional space. We also show its effectiveness in action clustering using K-medoids and diverse action sampling using manifold Precis [19]. Finally, the low dimensional features can easily be reconstructed back to the original manifold, enabling applications such as exploring and visualizing the *space of actions* in an intuitive manner.

Elastic representations for manifold sequences is relatively new and the lower dimensional embedding of such sequences has remained unexplored in computer vision. We employ the transport square-root velocity function (TSRVF) representation – a recent development in statistics [22], to provide a warp invariant representation to the action sequences. The TSRVF is also advantageous as it provides a functional representation that is Euclidean. Exploiting this we propose to learn the low dimensional embedding with a manifold functional variant of principal component analysis (mfPCA). In other words, we are interested in parameterization of sequences, i.e. for N actions $A_i(t), i = 1 \ldots N$ our goal is to learn $\mathcal{F}$ such that $\mathcal{F}(x) = A_i$ where $x \in \mathbb{R}^k$ is the set of parameters. Such a model will allow us to compare actions by simply comparing them in their parametric space with respect to $\mathcal{F}$, with significantly faster distance computations, while being able to reconstruct the original actions. In this work, we make the assumption that $\mathcal{F}$ is linear and learn it using mfPCA. In our experiments, we show that this is a suitable assumption for action recognition.

Broader impact: While one advantage of embedding action sequences into a lower dimensional space is low cost of storage and transmission, perhaps the biggest advantage is the reduction in complexity of search and retrieval in action spaces. Although this work concerns itself primarily with recognition and reconstruction, it is easy to see the opportunities these embeddings present in search applications given that the search space dimension is now $\sim 250\times$ smaller. We conclusively demonstrate that the embeddings are as discriminative as their original features, therefore guaranteeing an accurate and fast search.

Contributions

1. The extension of the TSRVF representation for human actions by modeling trajectories on the Grassmann manifold and the product space of $SE(3) \times .. \times SE(3)$.
2. The first embedding of Riemannian trajectories to lower dimensional spaces in a warp invariant manner which enables faster computations and lower storage.
3. The embedded features outperform many state-of-the-art approaches in action recognition on three benchmark datasets. Their effectiveness is also demonstrated in action clustering and diverse action sampling.
4. Easy reconstruction back to the original manifold enables the visual exploration of the latent variable style space of actions.

1.1. Related Work

Elastic metrics for trajectories: The TSRVF was recently introduced in statistics [22] as a way to represent trajectories on Riemannian manifolds such that the distance between two trajectories is invariant to identical time-warpings. The representation itself lies on a tangent space and is therefore Euclidean, this is discussed further in section 3. The representation was then applied to the problem of visual speech recognition by warping trajectories on the space of SPD matrices [23]. More recently, a vector space version of the representation known as the Square-Root Velocity Function (SRVF) was applied to skeletal action recognition with promising results [5]. We differentiate

our contribution as the first to use the TSRVF representation by representing actions as trajectories in high dimensional non-linear spaces. We use the skeletal feature recently proposed in [29], which models each skeleton as a point on the space of $SE(3) \times \cdots \times SE(3)$. Rate invariance for activities has been addressed before [27, 32], for example [27] models the space of all possible 'warpings' of an action sequence. Such techniques can align sequences correctly, even when features are multi-modal [32]. However, most of the techniques are used for recognition which can be achieved with a similarity measure, but we are interested in a representation which serves a more general purpose to 1) provide an effective metric for comparison, recognition, retrieval, etc. and 2) provide a framework for efficient lower dimensional coding which also enables recovery back to the original space.

Low dimensional data embedding Principal component analysis has been used extensively in statistics for dimensionality reduction of linear data. It has also been extended to model a wide variety of data types [12]. For high dimensional data in $\mathbb{R}^n$, manifold learning (or non-linear dimensionality reduction) techniques [24, 17] attempt to identify the underlying low dimensional manifold while preserving specific properties of the original space. Using a robust metric, one could ideally use such techniques for coding, but they do not provide a way of reconstructing the original manifold data. For data already lying on a known manifold, geometry aware mapping of SPD matrices [9] constructs a lower-dimensional SPD manifold, and principal geodesic analysis (PGA) [6] identifies the primary geodesics along which there is maximum variability of data points. We are interested in identifying the variability of sequences instead. The Gaussian process latent variable model (GPLVM) [13] and its variants, are a set of techniques that perform non-linear dimensionality reduction for data in $\mathbb{R}^N$, while allowing for reconstruction back to the original space. Further, these techniques deal with static points instead of sequences of trajectories, which is the primary concern of this work. Recently, dictionary learning methods for data lying on Riemannian manifolds have been proposed [11, 10] and could potentially be used to code sequential data but they can be expected to be computationally intensive. Since the TSRVF representation lies on a tangent space, we are able to employ traditional vector space techniques for significantly faster learning. Comparing actions in the latent variable space is similar in concept to learning a linear dynamical system [25] for Euclidean data, where different actions can be compared in the parametric space of the model.

2. Mathematical Preliminaries

In this section we will briefly introduce the properties of the product space $SE(3) \times \cdots \times SE(3)$ and the Grassmann manifold which are considered in this work. For an introduction to Riemannian manifolds, we refer the reader to [1, 3].

2.1. Product Space of the Special Euclidean Group

For action recognition, we represent a stick figure as a combination of relative transformations between joints, as proposed in [29]. The resulting feature for each skeleton is interpreted as a point on the product space of $SE(3) \times \cdots \times SE(3)$. The skeletal representation explicitly models the 3D geometric relationships between various body parts using rotations and translations in 3D space [29]. These transformation matrices lie on the curved space known as the Special Euclidean group $SE(3)$. Therefore the set of all transformations lies on the product space of $SE(3) \times \cdots \times SE(3)$.

The special Euclidean group, denoted by $SE(3)$ is a Lie group, containing the set of all 4×4 matrices of the form

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

where R denotes the rotation matrix, which is a point on the special orthogonal group $SO(3)$ and $\overrightarrow{d}$ denotes the translation vector, which lies in $\mathbb{R}^3$. The 4×4 identity matrix I_4 is an element of $SE(3)$ and is the identity element of the group. The exponential map, which is defined as $\exp_{SE(3)} : \mathfrak{se}(3) \to SE(3)$ and the inverse exponential map, defined as $\log_{SE(3)} : SE(3) \to \mathfrak{se}(3)$ are used to traverse between the manifold and the tangent space respectively. The exponential and inverse exponential maps for $SE(3)$ are simply the matrix exponential and matrix logarithms respectively, from the identity element I_4. For efficient implementations of a general exponential and inverse exponential maps between any two arbitrary points, we refer the reader to [15]. The tangent space at I_4 of a $SE(3)$ is called the Lie algebra of $SE(3)$, denoted by $\mathfrak{se}(3)$. It is a 6-dimensional space formed by matrices of the form:

$$B = \begin{bmatrix} U & \overrightarrow{w} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -u_3 & u_2 & w_1 \\ u_3 & 0 & -u1 & w_2 \\ -u_2 & u_1 & 0 & w_3 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2}$$

where U is a 3×3 skew-symmetric matrix and $\overrightarrow{w} \in \mathbb{R}^3$. An equivalent representation of B is $\text{vec}(B) = [u_1, u_2, u_3, w_1, w_2, w_3]$ which lies on $\mathbb{R}^6$.

These tools are trivially extended to the product space, for example the identity element of the product space is simply $(I_4, I_4, \ldots, I_4)$ and the Lie algebra is $\mathfrak{m} = \mathfrak{se}(3) \times \cdots \times \mathfrak{se}(3)$. Parallel transport on the product space is simply the parallel transport of the point on component spaces. Further, the $SE(3)$ is a semi-direct product space denoted by $SO(3) \rtimes \mathbb{R}^3$. Therefore, the parallel transport on $SE(3)$ is achieved by transporting a tangent vector in $SO(3)$ and

combining it with $\overrightarrow{d}$, which can be used as it is since it lies in $\mathbb{R}^3$. Let $T_{O_1}(SO(3))$ denote the tangent space at $O_1 \in SO(3)$, then the parallel transport, W_t, of a tangent $W \in T_{O_1}(SO(3))$ from O_1 to O_2 is given by the following: let $A = \texttt{logm}(O_1, O_2)$, and $O_t = \texttt{expm}(O_1, 0.5A)$, then $W_t = O_1 O_t (O_1^T W) O_t$.

2.2. Grassmann Manifold as a Shape Space

To visualize the action space, we also use shape silhouettes of an actor for different activities. These are interpreted as points on the Grassmann manifold. To obtain a shape as a point, we first obtain a landmark representation of the silhouette by uniformly sampling the shape. Let $L = [(x_1, y_1), (x_2, y_2) \ldots, (x_m, y_m)]$ be an $m \times 2$ matrix that defines m points on the silhouette whose centroid has been translated to zero. The affine shape space [7] is useful to remove small variations in camera locations or the pose of the subject. Affine transforms of a base shape L_{base} can be expressed as $L_{affine}(A) = L_{base} A^T$, and this multiplication by a full rank matrix on the right preserves the column-space of the matrix, L_{base}. Thus the 2D subspace of $\mathbb{R}^m$ spanned by the columns of the shape, i.e. $span(L_{base})$ is invariant to affine transforms of the shape. Subspaces such as these can be identified as points on a Grassmann manifold [26].

Denoted by, $\mathcal{G}_{k,m-k}$, the Grassmann manifold is the space whose points are k-dimensional hyperplanes (containing the origin) in $\mathbb{R}^m$. An equivalent definition of the Grassmann manifold is as follows: To each k-plane, ν in $\mathcal{G}_{k,m-k}$ corresponds a unique $m \times m$ orthogonal projection matrix, P which is idempotent and of rank k. If the columns of a tall $m \times k$ matrix Y spans ν then $YY^T = P$. Then the set of all possible projection matrices $\mathbb{P}$, is diffeomorphic to $\mathcal{G}$. The identity element of $\mathbb{P}$ is defined as $Q = diag(\mathbf{1}_k, \mathbf{0}_{m-k})$, where $\mathbf{1}_a, \mathbf{0}_b$ are vector of a 1s and b 0s respectively. The Grassmann manifold $\mathcal{G}$(or $\mathbb{P}$) is a quotient space of the orthogonal group, $O(m)$. Therefore, the geodesic on this manifold can be made explicit by lifting it to a particular geodesic in $O(m)$ [20]. Then the tangent, X, to the lifted geodesic curve in $O(m)$ defines the velocity associated with the curve in $\mathbb{P}$. The tangent space of $O(m)$ at identity is $o(m)$, the space of $m \times m$ skew-symmetric matrices, X. Moreover in $o(m)$, the Riemannian metric is just the inner product of $\langle X_1, X_2 \rangle$ = trace$(X_1 X_2^T)$ which is inherited by $\mathbb{P}$ as well.

The geodesics in $\mathbb{P}$ passing through the point Q (at time t = 0) are of the type $\alpha : (-\epsilon, \epsilon) \mapsto \mathbb{P}, \alpha(t) = \exp(tX)Q\exp(-tX)$, where X is a skew-symmetric matrix belonging to the set M where

$$M = \left\{ \begin{bmatrix} 0 & A \\ -A^T & 0 \end{bmatrix} : A \in \mathbb{R}^{k,n-k} \right\} \subset o(m) \qquad (3)$$

Therefore the geodesic between Q and any point P is completely specified by an $X \in M$ such that exp$(X)Q$exp$(-X) = P$. We can construct a geodesic between any two points $P_1, P_2 \in \mathbb{P}$ by rotating them to Q and some $P \in \mathbb{P}$. Readers are referred to [20] for more details on the exponential and logarithmic maps of $\mathcal{G}_{k,m-k}$.

3. Rate Invariant Sequence Comparison

In this section we describe the Transport Square Root Velocity Function (TSRVF), recently proposed in [22] as a representation to perform warp invariant comparison between multiple Riemannian trajectories. Using the TSRVF representation for human actions, we propose to learn the latent function space of these Riemannian trajectories in a much lower dimensional space. As we demonstrate in our experiments, such a mapping also provides some robustness to noise which is essential when dealing with noisy sensors.

Let α denote a smooth trajectory on $\mathcal{M}$ and let $\mathbb{M}$ denote the set of all such trajectories: $\mathbb{M} = \{\alpha : [0,1] \mapsto \mathcal{M} | , \alpha \text{ is smooth}\}$. Also define Γ to be the set of all orientation preserving diffeomorphisms of [0,1]: $\Gamma = \{\gamma \mapsto [0,1] | \gamma(0) = 0, \gamma(1) = 1, \gamma \text{ is a diffeomorphism}\}$. It is important to note that γ forms a group under the composition operation. If α is a trajectory on $\mathcal{M}$, then $\alpha \circ \gamma$ is a trajectory that follows the same sequence of points as α but at the evolution rate governed by γ. The TSRVF [22] for a smooth trajectory $\alpha \in \mathbb{M}$ is the parallel transport of a scaled velocity vector field of α to a reference point $c \in M$ according to:

$$h_\alpha(t) = \frac{\dot{\alpha}(t)_{\alpha(t) \mapsto c}}{\sqrt{|\dot{\alpha}(t)|}} \in T_c(\mathcal{M}) \qquad (4)$$

where $|\,.\,|$ denotes the norm related to the Riemannian metric on $\mathcal{M}$ and $T_c(\mathcal{M})$ denotes the tangent space of $\mathcal{M}$ at c. Since α is smooth, so is the vector field h_α.

The choice of reference point c needs to be selected in a consistent manner in this framework, and can potentially affect the results. This typically depends on the application and if most of the trajectories pass through or close to a common point, then the point becomes a natural candidate for c, for example the Riemannian center of mass (RCM) [8] can be used generally. For the product space $SE(3) \times \cdots \times SE(3)$ a natural candidate for c is the identity element $I_4 \times \cdots \times I_4$. In the case of landmarks on shapes, a possible candidate for c could be a neutral stance, since most actions can be seen as originating from or ending in a neutral stance.

Distance between TSRVFs: Since the TSRVFs lie on $T_c(\mathcal{M})$, the distance is measured by the standard $\mathbb{L}^2$ norm given by

$$d_h(h_{\alpha_1}, h_{\alpha_2}) = \left(\int_0^1 |h_{\alpha_1}(t) - h_{\alpha_2}(t)|^2 \right)^{\frac{1}{2}}. \qquad (5)$$

If a trajectory α is warped by γ, to result in $\alpha \circ \gamma$, the TSRVF of the warped trajectory is given by:

$$h_{\alpha\circ\gamma}(t) = h_\alpha(\gamma(t))\sqrt{\dot{\gamma}(t)} \tag{6}$$

The distance between TSRVFs remains unchanged to warping, i.e. $d_h(h_{\alpha_1}, h_{\alpha_2}) = d_h(h_{\alpha_1\circ\gamma}, h_{\alpha_2\circ\gamma})$. The invariance to group action is important as it allows us to compare two trajectories using the optimization problem stated next.

Metric invariant to temporal variability: Any two trajectories α_1, α_2 are said to be equivalent, $\alpha_1 \sim_1 \alpha_2$, if there exists a sequence $\{\gamma_k\} \in \Gamma$ such that $\lim_{k\mapsto\infty} h_{\alpha_1\circ\gamma_k} = h_{\alpha_2}$ this convergence is measured under the $\mathbb{L}^2$ metric. So if $\alpha_1 \sim_1 \alpha_2$, this means that $h_{\alpha_1} = h_{\alpha_2}$ and that in turn implies that $\alpha_1 \sim_2 \alpha_2$, where $\sim_2$ is an equivalence relation defined by $\alpha \sim_2 \beta$ if $h_\alpha = h_\beta$[22]. Let $\mathcal{H}/\sim_2$ be the corresponding quotient space, this can be bijectively identified with the set $\mathbb{M}/\sim_2$ using $[h_\alpha]_2 \mapsto [\alpha]_2$.

The distance d_s on $\mathcal{H}/\sim$ (or $\mathbb{M}/\sim$), where the subscript is dropped for convenience, is given by:

$$\begin{aligned} d_s([\alpha_1],[\alpha_2]) &\equiv \inf_{\gamma\in\Gamma} d_h(h_{\alpha_1}, h_{\alpha_2\circ\gamma}) \\ &= \inf_{\gamma\in\Gamma} \left(\int_0^1 \left| h_{\alpha_1}(t) - h_{\alpha_2}(\gamma(t))\sqrt{\dot{\gamma}(t)} \right|^2 dt \right)^{\frac{1}{2}} \end{aligned} \tag{7}$$

The minimization over Γ is solved for using dynamic programming. Here one samples the interval $[0, 1]$ using T discrete points and then restricts to only piecewise linear γ that passes through the $T \times T$ grid. By construction, this distance is invariant to warping, i.e. for $\gamma_1, \gamma_2 \in \Gamma, d_s([\alpha_1 \circ \gamma_1], [\alpha_2 \circ \gamma_2]) = d_s([\alpha_1], [\alpha_2])$.

3.1. The Latent Function Space of Actions

Typical representations for actions tend to be very high dimensional, and we propose to learn the low dimensional embedding. Note, this is the lower dimensional manifold of sequences which is different from the manifold that represents the individual features such as the shape silhouettes on the Grassmann, etc. We use principal component analysis (PCA) to learn the embedding, and show its effectiveness in recognition and reconstruction. The assumption that the lower dimensional manifolds are linear can be relaxed with more general dimensionality reduction methods such as the Gaussian Process Latent Variable Model (GPLVM) [13], manifold learning techniques such as Isomap [24], and LLE [17].

Manifold Functional PCA (mfPCA): The TSRVF representation allows us to study first and second order statistics on *entire sequences of actions* and enables us to define quantities such as the variability of actions, which we can exploit to perform PCA. We utilize the TSRVF to obtain the ideal warping between sequences, such that the warped sequence is equivalent to its TSRVF. To identify the principal components, we represent the sequences as deviations from a reference sequence using tangent vectors. For manifolds such as $SE(3)$ the natural "origin" I_4 can be used, in other cases the sequence Riemannian center of mass [22] by definition lies equi-distant from all the points and therefore is a suitable candidate. However in all our experiments, we found the tangent vectors obtained from the Riemannian center of mass-sequence to be much more robust and discriminative. Next, we obtain the shooting vectors, which are the tangent vectors one would travel along, starting from the average sequence $\mu(t)$ at $\tau = 0$ to reach the i^{th} action $\tilde{\alpha}_i(t)$ at time $\tau = 1$. Note here that τ is the time in the sequence space which is different from t, which is time in the original manifold space. We can use the fact that the sequence space is Euclidean and perform vector space PCA, this is outlined in algorithm 1. The combined shooting vectors can be understood as a *sequence tangent* that takes us from one point to another in sequence space, in unit time. These sequence tangents lie in $\mathbb{R}^N$ and therefore other coding schemes such as dictionary learning can be employed. Reconstructing back the features from the mfPCA space is

Algorithm 1 mfPCA - Manifold Functional PCA

1: **Input**: $\alpha_1(t), \alpha_2(t) \ldots \alpha_N(t)$
2: Compute Riemannian center of mass $\mu(t)$, which also aligns $\tilde{\alpha}_1(t), \tilde{\alpha}_2(t) \ldots \tilde{\alpha}_N(t)$ using TSRVF [22].
3: **for** $i \leftarrow [1 \ldots N]$ **do**
4: **for** $t \leftarrow [1 \ldots T]$ **do**
5: Compute shooting vectors $v(i,t) \in T_{\mu(t)}(M)$ as $v(i,t) = exp^{-1}_{\mu(t)}(\tilde{\alpha}_i(t))$
6: **end for**
7: Define $V(i) = [v(i,1)^T\ v(i,2)^T\ \ldots\ v(i,T)^T]^T$
8: **end for**
9: V is the feature matrix which is used to learn a principal basis, dictionary etc.

Algorithm 2 Reconstructing Non Euclidean Features from mfPCA

1: **Input**: Coordinates in mfPCA space, $C \in \mathbb{R}^{d\times N}, d << D$, $P \in \mathbb{R}^{D\times d}$, the orthogonal basis, $\mu(t)$, the average sequence.
2: **for** $i \leftarrow [1 \ldots N]$ **do**
3: Obtain shooting vectors $\widehat{V}_i = PC$
4: Rearrange $\widehat{V}_i$ as a $D = m \times T$ matrix, where T is the length of each sequence.
5: **for** $t \leftarrow [1 \ldots T]$ **do**
6: $\widehat{S}_i(t) = exp_{\mu(t)}(\widehat{V}_i(t), 1)$
7: **end for**
8: **end for**

shown in algorithm 2.

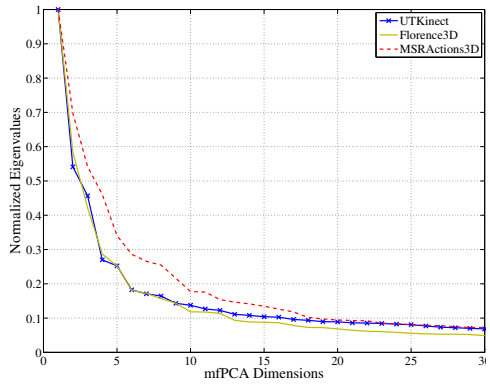


Figure 2: Eigenvalue decay for MSRActions3D [14], UTKinect [31], and Florence3D [18] datasets obtained with mfPCA. UTKinect and Florence3D have 10 and 9 different classes respectively, as a result the corresponding eigenvalue decay saturates at around 10 dimensions. MSRActions consists of 20 classes and the decay saturates later at around 20.

Eigenvalue decay with mfPCA: Figure 2 shows the eigenvalue decay after performing mfPCA on three commonly used datasets in skeletal action recognition. It is observed that mfPCA does a good job of approximating the different classes in the product space of $SE(3) \times \cdots \times SE(3)$. The MSRActions dataset [14] contains 20 classes and correspondingly the eigenvalue decay flattens around 20. In comparison the UTKinect [31] and Florence3D [18] datasets contain 10 and 9 classes of actions respectively, which is reflected in the eigenvalue decay that flattens closer to around 10. Since the TSRVF is a functional representation for sequential data, mfPCA is a generalization of functional PCA to manifolds. mfPCA is also a generalization of Principal Geodesic Analysis (PGA) [6] to trajectories. As is common, features in the mfPCA tend to be lower dimensional and more robust to noise, which is helpful in reducing the amount of pre/post processing required for optimal performance.

4. Experimental Evaluation

We perform recognition and reconstruction to demonstrate the utility of our low dimensional representation. For action recognition, we use a recently proposed feature called Lie algebra relative pairs (LARP) [29] for skeleton action recognition. This feature maps each skeleton to a point on the product space of $SE(3) \times SE(3) \cdots \times SE(3)$, where it is modeled using transformations between joint pairs. It was shown to be very effective on three benchmark datasets - UTKinect [31], Florence3D [18], and MSR Actions3D [14]. We show that using geometry aware warping results in significant improvement in recognition. Further, we show that it is possible to do so with a representational feature dimension that is $250\times$ smaller than state-of-the-art.

Florence3D actions dataset [18] contains 9 actions performed by 10 different subjects repeated two or three times by each actor. There are 15 joints on the skeleton data collected using the Kinect sensor. There are a total of 182 relative joint interactions which are encoded in the features.

UTKinect actions dataset [31] contains 10 actions performed by 10 subjects, each action is repeated twice by the actor. Totally, there are 199 action sequences. Information regarding 20 different joints is provided. There are a total of 342 relative joint interactions.

MSRActions3D dataset [14] contains a total of 557 labeled action sequences consisting of 20 actions performed by 10 subjects. There are 20 joint locations provided on the skeletal data, which gives 342 relative joint interactions.

UMD actions dataset [27]: This is a relatively constrained dataset, which has a static background allowing us to easily extract shape silhouettes. It contains 100 sequences consisting of 10 different actions repeated 10 times by the same actor. For this dataset, we use the shape silhouette of the actor as our feature, because of its easy visualization as compared to other non-linear features.

4.1. Alternative Representations

We compare the performance of our representation with various other recently proposed related methods:

Lie Algebra Relative Pairs (LARP): Recently proposed in [29], this feature is shown to model skeletons effectively. We will compare our results to those obtained using the LARP feature with warping obtained from DTW and unwarped sequences as baselines.

BP + SRVF : A skeleton is a collection of body parts where each skeletal sequence is represented as a combination of multiple body part sequences, proposed in [5]. It is also relevant to our work because the authors use the SRVF for ideal warping, which is the vector space version of the representation used in this paper. The representational dimension is calculated assuming the number of body parts $N_{Jp} = 10$, per skeleton[5].

Principal Geodesic Analysis (PGA)[6]: Performs PCA on the tangent space of static points on a manifold. We code individual points using this technique and concatenate the final feature vector before classification.

4.2. Evaluation Settings

The skeletal joint data obtained from low cost sensors are often noisy, as a result of which post-processing methods such as Fourier Temporal Pyramid (FTP) [30] have been shown to be very effective for recognition in the presence of noise. FTP is also a powerful tool to work around alignment issues, as it transforms a time series into the Fourier domain and discards the high frequency components. By the nature

Feature	Representational Dimension	Accuracy
BP+SRVF [5]	30,000	87.04
LARP [29]	38,220	86.27
LARP+DTW [29]	38,220	86.74
LARP+PGA [6]	6370	79.01
LARP+TSRVF	38200	89.50
LARP+mfPCA	**110**	**89.67**

Table 1: Recognition performance on the Florence3D actions dataset [18] for different feature spaces.

Feature	Representational Dimension	Accuracy
BP+SRVF [5]	60000	91.10
HOJ3D [31]	N/A	90.92
LARP [29]	151,848	93.57
LARP+DTW [29]	151,848	92.17
LARP+PGA [6]	25,308	91.26
LARP+TSRVF	151,848	94.47
LARP+mfPCA	**105**	**94.87**

Table 2: Recognition performance on the UTKinect actions dataset [31].

Feature	Representational Dimension	Accuracy
BP + SRVF [5]	60000	$\mathbf{87.28 \pm 2.99}$
HON4D [16]	N/A	82.15 ± 4.18
LARP[29]	155,952	75.57 ± 3.43
LARP+ DTW[29]	155,952	78.75 ± 3.08
LARP+PGA [6]	25,992	72.06 ± 3.12
LARP+TSRVF	155,952	84.62 ± 3.08
LARP+mfPCA	**250**	$\mathbf{85.16 \pm 3.13}$

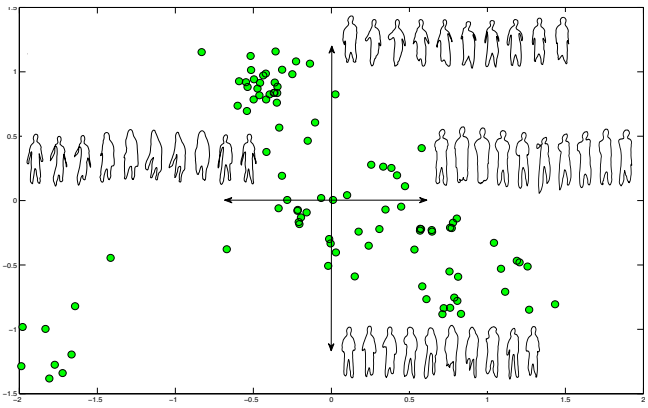
Table 3: Recognition performance on the MSRActions3D dataset [14] following the protocol of [16] by testing on 20 classes, with all possible combinations of test train subjects.

of FTP, the final feature is invariant to any form of warping. One of the contributions of this work is to demonstrate the effectiveness of geometry aware warping over conventional methods, and then explore the space of these warped sequences, which is not easily possible with FTP. Therefore, we perform our recognition experiments on the non-Euclidean features sequences without FTP. For classification, we use a one-vs-all SVM classifier following the protocol of [29], and set the C parameter to 1 in all our experiments. For the Florence3D and UTKinect datasets we use five different combinations of test-train scenarios and average the results. For the MSRActions dataset, we follow the train-test protocol of [16] by performing recognition on all 242 scenarios of 10 subjects of which half are used for training, and the rest for testing.

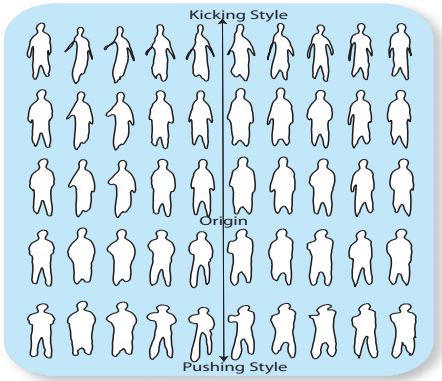
Recognition results The recognition rates for Florence 3D, UTKinect, and MSRActions3D are shown in tables 1, 2 and 3 respectively. It is clear from the results that using TSRVF on a Riemannian feature, leads to significant improvement in performance. Further, using mfPCA improves the results slightly, perhaps due to robustness to noise, but more importantly, reduces the representational dimension of each action by a factor of nearly 250. The improvements are significant compared to using DTW as a baseline; the performance is around 3% better on Florence3D, 2% on UTKinect, and 7% averaged over all test train variations on MSR Actions 3D. Although BP+SRVF [5] has higher recognition numbers on the MSRActions3D, our contribution lies in the significant advantage obtained using the LARP features with mfPCA (over 7% on average). We observed that simple features in $\mathbb{R}^N$ performed exceedingly well on MSRActions3D, for example using relative joint positions (given by $\overrightarrow{v} = J_1 - J_2$, where J_1 and J_2 are 3D coordinates joints 1 and 2.) on the MSRActions3D with SRVF and PCA we obtain $87.17 \pm 3.08\%$ by embedding every action into $\mathbb{R}^{250}\times$, which is similar to [5], but in a much lower dimensional space. We also show that performing PCA on the "sequence tangents" is significantly better than performing PCA on individual time samples using Principal Geodesic Analysis. The dimensions for LARP features are calculated as $6 \times J \times T$, where J is the number of relative joint pairs per skeleton, and T is the number of frames per video. We learn the mfPCA basis using training data, and project the test data onto the orthogonal basis.

Reconstruction results: Once we have mapped the actions onto their lower dimensional space using mfPCA, we can reconstruct them back easily using algorithm 2. We show that high dimensional action sequences that lie in non-Euclidean spaces can be effectively embedded into a lower dimensional latent variable space. Figure 3a, shows the space of 100 actions from the UMD dataset mapped onto $\mathbb{R}^2$. Note, that since we are only using 2 dimensions, there is a loss of information, but the variations are still visually discernible. The primary components in this dataset are shown on the two perpendicular axes and figure 3b shows the sampling of one axis at different points. As expected, the "origin" of the dataset contains no information about any action, but moving in the positive or negative direction of the axis results in different *styles* as shown.

Diverse sequence sampling with mfPCA: To further demonstrate the generalizability of the lower dimensional features to other sequence based algorithms, we use K-medoids to cluster actions and Precis [19] for diverse action sampling. In the experiment on the UMD actions dataset, we constructed a collection of actions that were chosen such that different classes had significantly different populations

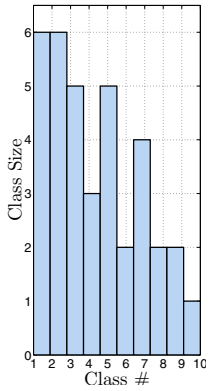


(a) Exploring the action space in $\mathbb{R}^2$.

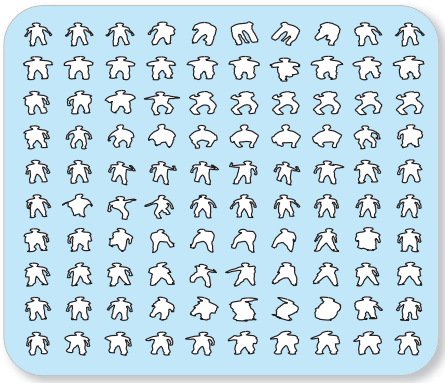


(b) Sampling along a principal component in the mfPCA space.

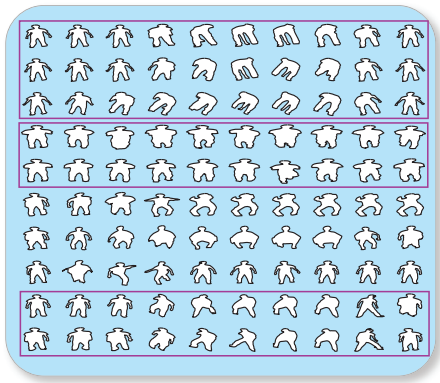
Figure 3: Exploring the latent variable space of actions in the UMD actions dataset using mfPCA. Notice in fig 3b, the "origin" contains no information about any action, and moving along an axis provides different abstract style information.



(a) A dataset with skewed class proportions



(b) Row-wise: Actions sampled by functional-Precis



(c) Row-wise: Centers obtained using functional K-medoids

Figure 4: Diverse action sampling using Precis [19] by sampling in mfPCA space $\in \mathbb{R}^{10}$. K-medoids picks more samples (marked) from classes that have a higher representation which are actions #1,#2 and #7 here. The exemplar selection is performed $\sim 500\times$ faster in the mfPCA space.

in the collection, a distribution of the action classes is shown in figure 4a. Action centers obtained with K-medoids is shown in figure 4c and as expected classes which have a higher population are over represented in the chosen samples as compared to Precis (figure 4b) which is invariant to the distribution. Due to the low dimensional Euclidean representation, these techniques can be easily extended to suit sequential data without any increased demand for computational resources.

5. Conclusion & Future Work

In this paper we introduced techniques to explore and analyze sequential data on Riemannian manifolds, applied to human actions. We employ the TSRVF space [22], which provides an elastic metric between two trajectories on a manifold, to learn the latent variable space of actions. We demonstrate these ideas on the curved product space $SE(3) \times \cdots \times SE(3)$ for skeletal actions and the Grassmann manifold for shape silhouettes. We propose mfPCA which generalizes functional PCA to manifolds and PGA to sequences. The learned mfPCA dimension not only provides a compact and robust representation that outperforms many state of the art methods, but also the visualization of actions due to its ability to reconstruct original non-linear features. The proposed representation opens up several opportunities in fast search and retrieval of actions sequences from large databases which can largely benefit the computer vision community. An extension of this work could include utilizing better coding techniques such as sparse representations for even higher reduction in required data, while providing better reconstruction quality.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. 3

[2] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994. 2

[3] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry. Revised 2nd Ed.* Academic, New York, 2003. 3

[4] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *CVPR*, June 2009. 1

[5] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *Cybernetics, IEEE Transactions on*, PP(99):1–1, September 2014. 2, 6, 7

[6] P. T. Fletcher, C. Lu, S. M. Pizer, and S. C. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, August 2004. 1, 3, 6, 7

[7] C. R. Goodall and K. V. Mardia. Projective shape analysis. *Journal of Computational and Graphical Statistics*, 8(2), 1999. 4

[8] K. Grove and H. Karcher. How to conjugate C^1-close group actions. *Math.Z*, 132:11–20, 1973. 4

[9] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: Geometry-aware dimensionality reduction for SPD matrices. In *ECCV 2014*, pages 17–32, 2014. 3

[10] M. T. Harandi, C. Sanderson, C. Shen, and B. C. Lovell. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *ICCV*, pages 3120–3127, 2013. 3

[11] J. Ho, Y. Xie, and B. C. Vemuri. On a nonlinear generalization of sparse coding and dictionary learning. In *ICML (3)*, pages 1480–1488, 2013. 3

[12] I. Jolliffe. *Principal Component Analysis*. John Wiley & Sons, Ltd, 2005. 3

[13] N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16:329–336, 2004. 3, 5

[14] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010*, pages 9–14. IEEE, 2010. 2, 6, 7

[15] R. M. Murray, Z. Li, and S. S. Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 1994. 3

[16] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *(CVPR), 2013*, pages 716–723. IEEE, 2013. 7

[17] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000. 3, 5

[18] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013*, pages 479–485, 2013. 2, 6, 7

[19] N. Shroff, P. K. Turaga, and R. Chellappa. Manifold precis: An annealing technique for diverse sampling of manifolds. In *NIPS*, 2011. 2, 7, 8

[20] A. Srivasatava and E. Klassen. Bayesian geometric subspace tracking. *Advances in Applied Probability*, 36(1):43–56, March 2004. 4

[21] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu. Statistical shape analysis: Clustering, learning, and testing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(4), 2005. 1

[22] J. Su, S. Kurtek, E. Klassen, and A. Srivastava. Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking, and video surveillance. *Annals of Applied Statistics*, 8(1), 2014. 2, 4, 5, 8

[23] J. Su, A. Srivastava, F. D. M. de Souza, and S. Sarkar. Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition. In *CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 620–627, 2014. 2

[24] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 3, 5

[25] P. K. Turaga and R. Chellappa. Locally time-invariant models of human activities using trajectories on the Grassmannian. In *CVPR*, pages 2435–2441, 2009. 1, 2, 3

[26] P. K. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(11):2273–2286, 2011. 4

[27] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury. The function space of an activity. *IEEE CVPR*, pages 959–968, 2006. 3, 6

[28] A. Veeraraghavan, A. K. R. Chowdhury, and R. Chellappa. Matching shape sequences in video with applications in human movement analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1896–1909, 2005. 1

[29] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *(CVPR), 2014*, pages 588–595, June 2014. 1, 2, 3, 6, 7

[30] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *(CVPR), 2012*, pages 1290–1297, June 2012. 6

[31] L. Xia, C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW)2012*, pages 20–27. IEEE, 2012. 2, 6, 7

[32] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *(CVPR), 2012*, pages 1282–1289. IEEE, 2012. 3