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

Rushil Anirudh,[1,2] Pavan Turaga,[2,1] Jingyong Su,[3] Anuj Srivastava[4]
[1]School of Electrical, Computer, and Energy Engineering, Arizona State University. [2]School of Arts, Media, and Engineering, Arizona State University.
[3]Department of Mathematics & Statistics, Texas Tech University. [4]Department of Statistics, Florida State University.

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. 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. Further, 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 (see fig 1). 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). However, DTW is not a proper metric and does not naturally allow the estimation of statistical measures such as mean and variance of action trajectories. Instead we use the Transport Square-Root Velocity Function (TSRVF) [1], to provide a warp invariant representation to the action sequences.

Let α denote a smooth trajectory on a manifold, $\mathcal{M}$ 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]: $\Gamma = \{\gamma \mapsto [0,1] | \gamma(0) = 0, \gamma(1) = 1, \gamma \text{ is a diffeomorphism}\}$. 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 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:

$$h_\alpha(t) = \frac{\dot{\alpha(t)}_{\alpha(t) \mapsto c}}{\sqrt{|\dot{\alpha(t)}|}} \in T_c(\mathcal{M}) \tag{1}$$

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.

Manifold Functional PCA (mfPCA): We utilize the TSRVF to obtain the ideal warping between sequences, such that the warped sequence is equivalent to its TSRVF. This 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. To identify the principal components, we represent the sequences as deviations from a reference sequence using 'shooting' vectors. We can use the fact that the sequence space is Euclidean and perform vector space PCA. 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. Since PCA allows us to reconstruct back, we are able to visualize the latent space of actions, as shown in fig 2.

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

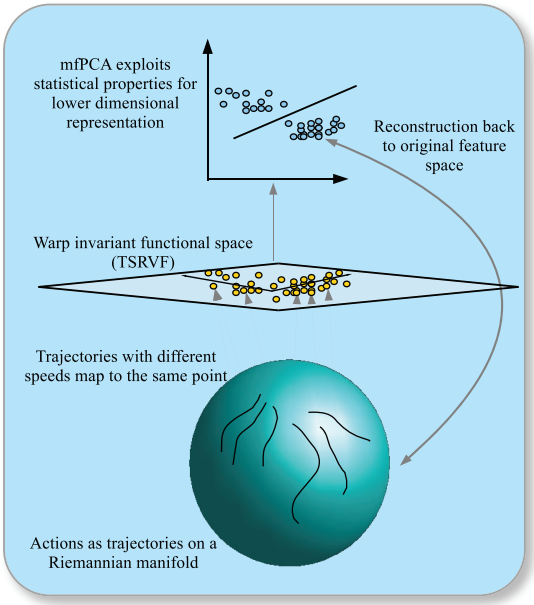

Figure 1: Overview of our work

$\mathcal{F}$ is linear and learn it using mfPCA. In our experiments, we show that this is a suitable assumption for action recognition.

We consider two types of features for human actions which lie on different manifolds - shape silhouettes on the Grassmann manifold [2] and skeletal joints as points on the product space $SE(3) \times \cdots \times SE(3)$[3]. We show that the lower dimensional embedding can accurately recognize actions on benchmark datasets, as well as the original features, on a significantly lower dimensional space.

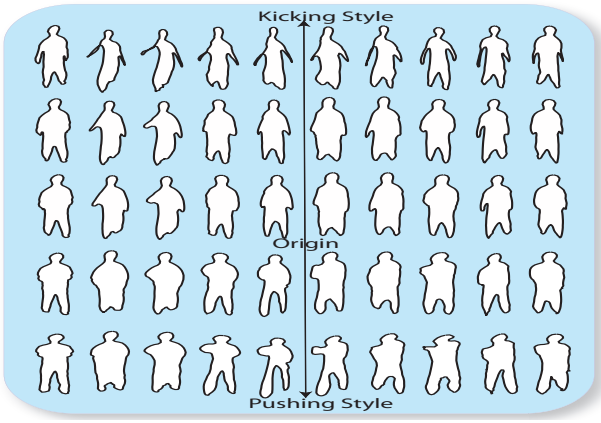

Figure 2: Exploring the latent variable space of actions in the UMD actions dataset using mfPCA. Notice the "origin" contains no information about any action, and moving along an axis provides different abstract style information.

[1] Jingyong Su, Sebastian Kurtek, Eric Klassen, and Anuj Srivastava. Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking, and video surveillance. *Annals of Applied Statistics*, 8(1), 2014.

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

[3] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *(CVPR), 2014*, pages 588–595, June 2014.