Rate-Invariant Analysis of Trajectories on Riemannian Manifolds with Application in Visual Speech Recognition

Jingyong Su
Department of Mathematics & Statistics
Texas Tech University, Lubbock, TX
jingyong.su@ttu.edu

Anuj Srivastava
Department of Statistics
Florida State University, Tallahassee, FL
anuj@stat.fsu.edu

Fillipe D. M. de Souza, Sudeep Sarkar
Department of Computer Science and Engineering
University of South Florida, Tampa, FL
fillipe@mail.usf.edu, sarkar@cse.usf.edu

Abstract

In statistical analysis of video sequences for speech recognition, and more generally activity recognition, it is natural to treat temporal evolutions of features as trajectories on Riemannian manifolds. However, different evolution patterns result in arbitrary parameterizations of these trajectories. We investigate a recent framework from statistics literature [15] that handles this nuisance variability using a cost function/distance for temporal registration and statistical summarization & modeling of trajectories. It is based on a mathematical representation of trajectories, termed transported square-root vector field (TSRVF), and the $\mathbb{L}^2$ norm on the space of TSRVFs. We apply this framework to the problem of speech recognition using both audio and visual components. In each case, we extract features, form trajectories on corresponding manifolds, and compute parametrization-invariant distances using TSRVFs for speech classification. On the OuluVS database the classification performance under metric increases significantly, by nearly 100% under both modalities and for all choices of features. We obtained speaker-dependent classification rate of 70% and 96% for visual and audio components, respectively.

1. Introduction

In this paper we focus on problems that deal with registration, comparison and summarization of trajectories on nonlinear manifolds. An important issue in the analyses is that trajectories are often not observed at standard times but, in fact, at random times. One motivation of this problem comes from activity recognition where features extracted from video frames are naturally represented as elements of nonlinear manifolds, and the temporal evolutions of activities can be treated as trajectories on those manifolds. For example, if one restricts to shapes of silhouettes in images, then an activity is a parametrized path on the shape space of contours [18]. However, as highlighted in [18, 21], the execution rate of activities is often arbitrary and that results in random parameterizations of corresponding trajectories. This parameterization variability results in artificially inflated distances between trajectories, even those within the same class, and distorts computations of summary statistics such as mean trajectory and variance along trajectories. Thus, temporal alignment and removal of temporal variability are important in statistical modeling of activities.

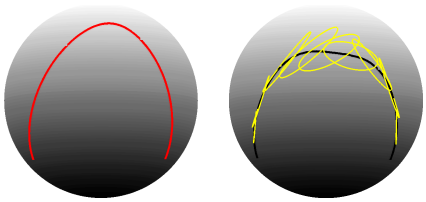


Figure 1. Left: a trajectory on $\mathbb{S}^2$; Right: mean and point-wise variance of this trajectory observed under arbitrary parameterizations.

The need for warping/alignment while analyzing trajectories can be motivated with a simple example. Consider the trajectory on a unit sphere shown in the left part of Fig. 1. We simulate a set of random evolution rates along this trajectory and generate several observations of this trajectory at these random times. These simulated trajectories are identical in terms of the points traversed but their evolutions, or parameterizations are quite different. The results

of cross-sectional mean and variance are shown in the right. We draw the sample mean trajectory in black and the sample variance at discrete times using tangential ellipses. Not only is the mean quite different from the original curve, the variance is artificially exaggerated due to randomness in observation times. If we have observed the trajectory at fixed, synchronized times, then this problem will not exist.

In particular, we are interested in the problem of speech classification using close-up videos of human facial movements. Speech classification is important because it allows computers to interpret human speech and take appropriate actions. The goal here is to develop efficient algorithms that input speech data and provide human-like interpretations. Applications of speech recognition can be found in systems for helping hearing-impaired, biometric security, human-machine interactions, manufacturing, home security systems, and so on. It should be emphasized that speech recognition can be performed with multiple modalities – the common speech data consists of both audio and visual components. In case the audio information is either not available or is corrupted by noise, it becomes important to understand the speech using only the visual data. This leads to the problem of visual speech recognition (VSR), which is also known as automatic lipreading. Previous work on VSR can be divided into two categories: sequential inference approaches and spatial-temporal descriptor approaches. Sequential inference approaches extract features from frames and recognize the sequence of features via state-based sequential inference models. Hidden Markov model (HMM) has often been used to perform VSR, e.g. [12]. Spatial-temporal descriptor approaches incorporate the temporal consistency among subsequent frames to capture the dynamics. These include the temporal derivatives [4] and local spatial-temporal descriptors (STD) [19].

The process of visual speech recognition is to understand the words uttered by speakers, derived from the visual cues. Movements of the tongue, lips, jaw, and other speech related articulators are involved to make sound. Speech is therefore a dynamic process involving these articulators. However, different speakers have varying appearances of their articulations due to different pronunciation habits. Even the same speaker can speak at different speeds in different situations. In addition, the imaging environments greatly affect the pose of producers, illumination and the image quality of the sequence to be analyzed. As a result, the overall appearance of two instances of the same utterance may vary considerably, though the intrinsic dynamics of movement are generally similar.

The theoretical framework used here has been introduced recently in statistics literature [15] but our goal here is to explore its applicability for computer vision problems, specifically in visual speech recognition. This is the first demonstration of the use of this mathematical theory for a computer vision problem. Our goal is to address temporal alignment of two sequences and not to find the best features. We do not claim that the particular features used in the algorithms here are optimal for classification. The temporal alignment would be useful irrespective of the choice of features. The paper is organized as follows. In Section 2, we summarize the general mathematical framework proposed in [15] for registration and comparison of trajectories. In Section 3, algorithms for computing Karcher mean and variance are provided. In Section 4, features for audio-visual speech recognition and their underlying spaces are introduced. In Section 5, experimental results involving OuluVS data are presented.

2. Mathematical framework

This material has been presented in a statistics paper earlier but is repeated here for convenience. Further details may be found in [15].

Let α denote a smooth trajectory on a Riemannian manifold of interest M, where M is endowed with a Riemannian metric $\langle \cdot, \cdot \rangle$. Let $\mathcal{M}$ denote the set of all such trajectories: $\mathcal{M} = \{\alpha : [0,1] \to M | \alpha \text{ is smooth}\}$. Also, define Γ to be the set of all diffeomorphisms of $[0,1]$: $\Gamma = \{\gamma : [0,1] \to [0,1] | \gamma(0) = 0, \gamma(1) = 1, \gamma \text{ is a diffeomorphism}\}$ with the ending points are preserved. It is important to note that Γ forms a group under the composition operation. If α is a trajectory on 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. More technically, the group Γ acts on $\mathcal{M}$, $\mathcal{M} \times \Gamma \to \mathcal{M}$, according to $(\alpha, \gamma) = \alpha \circ \gamma$. Given two smooth trajectories $\alpha_1,\ \alpha_2 \in \mathcal{M}$, we want to register points along the trajectories and compute a time-warping invariant distance between them. For performing comparison of trajectories, we need a metric and, at first, we consider a more conventional solution. Since M is a Riemannian manifold, we have a natural geodesic distance d_m between points on M. Using d_m, one can compare any two trajectories: $\alpha_1, \alpha_2 : [0,1] \to M$, as

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

Although this quantity represents a natural extension of d_m from M to $\mathcal{M}$, it suffers from the problem that $d_x(\alpha_1, \alpha_2) \neq d_x(\alpha_1 \circ \gamma_1, \alpha_2 \circ \gamma_2)$ in general. It is not preserved even when the same γ is applied to both the trajectories, i.e. $d_x(\alpha_1, \alpha_2) \neq d_x(\alpha_1 \circ \gamma, \alpha_2 \circ \gamma)$ generally. If we have an equality in the last case, for all γ's, then one can develop a fully invariant distance and use it to register trajectories properly, as described later. So, the failure to have this equality is in fact a key issue that forces us to look for other solutions in situations where trajectories are observed at random temporal evolutions.

2.1. Previous work

The fact that M is a Riemannian manifold presents a formidable challenge in developing a comprehensive framework. But this is not the only challenge. To clarify this part, let us consider this question: How has this registration and comparison problem been handled for trajectories in Euclidean spaces? In case $M = \mathbb{R}$, i.e. one is interested in registration and modeling of real-valued functions under random time-warpings, the problem has been studied by many authors, including [16]. [13] proposed a solution that applies to curves in arbitrary $\mathbb{R}^n$. One can also borrow solutions from problems in image registration where 2D and 3D images are registered to each other using a spatial warping instead of a temporal warping (see e.g. LDDMM technique [1]). A majority of the existing methods in Euclidean spaces formulate an objective function of the type: $\min_\gamma \left(\int_0^1 |\alpha_1(t) - \alpha_2(\gamma(t))|^2 dt + \lambda \mathcal{R}(\gamma) \right)$, where $|\cdot|$ is the Euclidean norm, $\mathcal{R}$ is a regularization term on the warping function γ, and $\lambda > 0$ is a constant. In case of a Riemannian manifold, one can modify the first term to obtain:

$$\min_\gamma \left(\int_0^1 d_m(\alpha_1(t), \alpha_2(\gamma(t)))^2 dt + \lambda \mathcal{R}(\gamma) \right) . \quad (2)$$

The main problem with this procedure is that: (a) it is not symmetric, i.e. the registration of α_1 to α_2 is not the same as that of α_2 to α_1, as pointed out by [3] and others, and (b) the minimum value is not a proper distance, so it cannot be used to compare trajectories. This sums up the fundamental dilemma in trajectory analysis – Eqn. 1 provides a metric between trajectories but does not perform registration and Eqn. 2 performs registration but is not a metric!

2.2. Rate-invariant comparison and registration

We introduce a representation of trajectories that will be used to compare and register them. We will assume that for any two points $p, q \in M$, we have an expression for parallel transporting any vector $v \in T_p(M)$ along the geodesic from p to q, denoted by $(v)_{p \to q}$. As long as p and q do not fall in the cut loci of each other, the geodesic between them is unique and the parallel transport is well defined. The measure of the set of cut locus on the manifolds of our interest is typically zero. So, the practical implications of this limitation are negligible. Let c be a point in M that we will designate as a reference point. We will assume that none of the observed trajectories pass through the cut locus of c to avoid the problem mentioned above.

Definition 1 *For any smooth trajectory $\alpha \in \mathcal{M}$, the transported square-root vector field (TSRVF) is a parallel transport of a scaled velocity vector field of α to a reference point $c \in M$ according to:*

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

where $|\cdot|$ is define by the Riemannian metric on M and the tangent space at c is denoted by $T_c(M)$, as shown in the left of Fig.2.

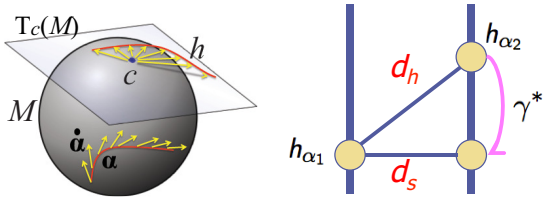


Figure 2. Left: TSRVF; Right: induced metric d_s.

Since α is smooth, so is the vector field h_α. Let $\mathcal{H} = \{h_\alpha | \alpha \in \mathcal{M}\}$ be the set of smooth curves in $T_c(M)$ obtained as TSRVFs of trajectories in M. If $M = \mathbb{R}^n$ with the Euclidean metric then h is exactly the square-root velocity function defined in [13].

The choice of reference point c used in Definition 1 is important in this framework and, in principle, will affect distances. But our experiments suggest that the results of registration and classification are quite stable with respect to this choice. An example is shown later in Fig. 4. Another remark is that instead of parallel transportation of scaled velocity vectors along geodesics, one can translate them along trajectories themselves, as was done in [7], but that requires c to be a common point of all trajectories.

Since a TSRVF is a path in $T_c(M)$, one can use the $\mathbb{L}^2$ norm to compare such trajectories.

Definition 2 *Let α_1 and α_2 be two smooth trajectories on M and let h_{α_1} and h_{α_2} be the corresponding TSRVFs. The distance between them is:*

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

The distance d_h, being the standard $\mathbb{L}^2$ norm, satisfies symmetry, positive definiteness, and triangle inequality. The main motivation of this setup – the TSRVF representation and $\mathbb{L}^2$ norm – comes from the following fact.

Theorem 1 *For any $\alpha_1, \alpha_2 \in \mathcal{M}$ and $\gamma \in \Gamma$, the distance d_h satisfies $d_h(h_{\alpha_1 \circ \gamma}, h_{\alpha_2 \circ \gamma}) = d_h(h_{\alpha_1}, h_{\alpha_2})$. In geometric terms, this implies that the action of Γ on $\mathcal{H}$ under the $\mathbb{L}^2$ metric is by isometries.*

It can be proved easily by plugging in the expression of $h_{\alpha \circ \gamma}$ and changing variables. Next we define a quantity that can be used as a distance between trajectories while being invariant to their temporal evolutions. To set up this definition, we first introduce an equivalence relation between trajectories. For any two trajectories α_1 and α_2, we define them to be equivalent, $\alpha_1 \sim \alpha_2$, if they have the same starting point and the TSRVF of one can be time-warped into

the TSRVF of the other using a sequence of warpings. It can be easily checked that $\sim$ forms an equivalence relation on $\mathcal{H}$ (and correspondingly $\mathcal{M}$).

Since we want our distance to be invariant to time-warpings of trajectories, we wish to compare trajectories by comparing their equivalence classes. Thus, our next step is to inherit the distance d_h to the set of such equivalence classes. Towards this goal, we introduce a set $\tilde{\Gamma}$ as the set of all absolutely continuous, non-decreasing functions $\gamma : [0,1] \to [0,1]$ such that $\gamma(0) = 0$ and $\gamma(1) = 1$. This set $\tilde{\Gamma}$ is a semigroup with the composition operation (it is not a group because the elements do not have inverses). The group Γ is a subset of $\tilde{\Gamma}$. The elements of $\tilde{\Gamma}$ warp the time axis of trajectories in M in the same way as elements of Γ, except they allow certain singularities. For a TSRVF $h_\alpha \in \mathcal{H}$, its equivalence class, or *orbit* under $\tilde{\Gamma}$, is given by $[h_\alpha] = \{h_{\alpha\circ\gamma} | h_\alpha \in \mathcal{H},\ \gamma \in \tilde{\Gamma}\}$. It can be shown that the orbits under $\tilde{\Gamma}$ are exactly the same as the closures of the orbits of Γ, defined as $[h_\alpha]_0 = \{h_{\alpha\circ\gamma} | \gamma \in \Gamma\}$, as long as α has non-vanishing derivatives almost everywhere. (The last condition is not restrictive since we can always re-parameterize α by the arc-length.) The closure is with respect to the $\mathbb{L}^2$ metric on $\mathcal{H}$.

Now we are ready to define the quantity that will serve as both the cost function for registration and the distance for comparison. This quantity is essentially d_h measured between not the individual trajectories but their equivalence classes, as shown in the right of Fig. 2.

Definition 3 *The distance d_s on $\mathcal{H}/\sim$ (or $\mathcal{M}/\sim$) is the shortest d_h distance between equivalence classes in $\mathcal{H}$:*

$$d_s([h_{\alpha_1}], [h_{\alpha_2}]) = \inf_{\gamma_1,\gamma_2 \in \tilde{\Gamma}} d_h(h_{\alpha_1\circ\gamma_1}, h_{\alpha_2\circ\gamma_2}) \,. \tag{3}$$

Now, since Γ is dense in $\tilde{\Gamma}$, for any $\delta > 0$, there exists a γ^* such that:

$$|d_h(h_{\alpha_1}, h_{\alpha_2\circ\gamma^*}) - d_s([h_{\alpha_1}], [h_{\alpha_2}])| < \delta \,.$$

This γ^* may not be unique but any such γ^* is sufficient for our purpose. Furthermore, since $\gamma^* \in \Gamma$, it has an inverse that can be used in further analysis.

Since the action of Γ is by isometries (Theorem 1) and the equivalence classes form closed sets, it can be shown that d_s is a proper distance, i.e. it satisfies symmetry, positive definiteness and triangle inequality, on the set $\mathcal{H}/\sim$. Additionally, it satisfies an important invariant property. For any $\gamma_1, \gamma_2 \in \Gamma$, we have:

$$d_s([h_{\alpha_1\circ\gamma_1}], [h_{\alpha_2\circ\gamma_2}]) = d_s([h_{\alpha_1}], [h_{\alpha_2}]).$$

In calculation of d_s between any two paths, we need to solve for the optimal correspondence between them according to:

$$\gamma^* = \operatorname*{argmin}_{\gamma\in\Gamma} \left(\int_0^1 |h_{\alpha_1}(t) - h_{\alpha_2}(\gamma(t))\sqrt{\dot{\gamma}(t)}|^2 dt \right)^{\frac{1}{2}} . \tag{4}$$

The minimization over Γ in Eqn. 4 can be solved in practice using the dynamic programming (DP) algorithm [2]. Here one samples the interval $[0,1]$ using T discrete points and then restricts to only piecewise linear γ that passes through that $T \times T$ grid. The search for the optimal trajectory on this grid is accomplished in $O(T^2)$ steps. $T = 50$ is used in this paper. While it is possible that the optimal mapping γ^* lies on the boundary of Γ, the DP algorithm provides an approximation using a piecewise linear map on a finite grid.

If we compare Eqn. 4 with Eqn. 2, we can immediately see the advantages of the proposed framework. Both equations present a registration problem between α_1 and α_2, but only the minimum value resulted from Eqn. 4 is a proper distance. Also, the optimal registration in Eqn. 4 remains the same if we change the order of the input functions. That is, the registration process is inverse consistent!

3. Summarization/Registration of multiple trajectories

An additional advantage of this framework is that one can compute an average of multiple trajectories and use it as a *template* for classification. Furthermore, this template can, in turn, be used for registering multiple trajectories. As stated before, this material has been presented in a statistics paper earlier but is repeated here for convenience. We will use the notion of the Karcher mean to define and compute average trajectories. Given a set of sample trajectories $\alpha_1, ..., \alpha_n$ on M, we represent them using the corresponding pairs $(\alpha_1(0), h_{\alpha_1}), (\alpha_2(0), h_{\alpha_2}), \ldots, (\alpha_n(0), h_{\alpha_n})$. We will compute the Karcher means of each component in their respective spaces: (1) the Karcher mean of $\alpha_i(0)$s are computed with respect to d_m in M, and (2) the Karcher mean of h_{α_i}s are with respect to d_s in $\mathcal{H}/\sim$. The latter Karcher mean is defined by: $h_\mu = \operatorname{argmin}_{[h_\alpha]\in\mathcal{H}/\sim} \sum_{i=1}^n d_s([h_\alpha], [h_{\alpha_i}])^2$. Note that $[h_\mu]$ is actually an equivalence class of trajectories and one can select any element of this mean class to help aligning multiple trajectories. The standard algorithm to compute the Karcher mean in [8, 15] is adapted to this problem as follows:

Algorithm 1 Karcher mean of multiple trajectories:
Compute the Karcher mean of $\{\alpha_i(0)\}s$ and set it to be $\mu(0)$.

1. *Initialization step: Select μ to be one of the original trajectories and compute its TSRVF h_μ.*
2. *Align each h_{α_i}, $i = 1, ..., n$, to h_μ according to Eqn. 4. That is, solve for γ_i^* using the DP algorithm and set $\tilde{\alpha}_i = \alpha_i \circ \gamma_i^*$.*
3. *Compute TSRVFs of the warped trajectories, $h_{\tilde{\alpha}_i}$, $i = 1, 2, \ldots, n$, and update h_μ as a curve in $T_c(M)$ according to: $h_\mu(t) = \frac{1}{n}\sum_{i=1}^n h_{\tilde{\alpha}_i}(t)$.*

4. *Define μ to be the integral curve associated with a time-varying vector field on M generated using h_μ, i.e. $\frac{d\mu(t)}{dt} = |h_\mu(t)|(h_\mu(t))_{c\to\mu(t)}$, and the initial condition $\mu(0)$.*

5. *Compute $E = \sum_{i=1}^{n} d_s([h_\mu],[h_{\alpha_i}])^2 = \sum_{i=1}^{n} d_h(h_\mu, h_{\tilde{\alpha}_i})^2$ and check it for convergence. If not converged, return to step 2.*

This algorithm provides two sets of outputs: an average trajectory denoted by the final μ and the set of aligned trajectories $\{\tilde{\alpha}_i\}$s. Therefore, this actually solves the problem of aligning multiple trajectories too. For each aligned trajectory $\tilde{\alpha}_i(t)$ at time t, the vector $v_i(t) \in T_{\mu(t)}(M)$ is computed such that a geodesic that goes from $\mu(t)$ to $\tilde{\alpha}_i(t)$ in unit time has the initial velocity $v_i(t)$. This is also called the *shooting vector* from $\mu(t)$ to $\tilde{\alpha}_i(t)$. Let $\hat{K}(t)$ be the sample covariance matrix of all the shooting vectors from $\mu(t)$ to $\tilde{\alpha}_i(t)$. The sample Karcher covariance at time t is given by $\hat{K}(t) = \frac{1}{n-1}\sum_{i=1}^{n} v_i(t)v_i(t)^T$, with the trace $\hat{\rho}(t) = \text{trace}(\hat{K}(t))$. This $\hat{\rho}(t)$ represents a quantification of the cross-sectional variance, as a function of t, and can be used to study the level of alignment of trajectories. As a simple illustration, we randomly simulate γs and apply them to a trajectory of 3×3 covariance matrices. For display, each covariance matrix is depicted using an ellipsoid. The simulated trajectories and γ's are shown in Fig. 3 (a) and (b). Then we compute the sample mean in two cases: without registration and with registration. The mean after registration in Fig. 3 (d) is a much better representation of data because the simulated trajectories are in fact the same, while the mean without registration in (c) losses the pattern of data. The comparison of $\hat{\rho}$'s in the two cases before and after are shown in Fig. 3 (e). The extraneous temporal variability is removed due to the registration. In addition, the sample mean is used to register five trajectories in the data and it results in the same trajectories as the mean in (d).

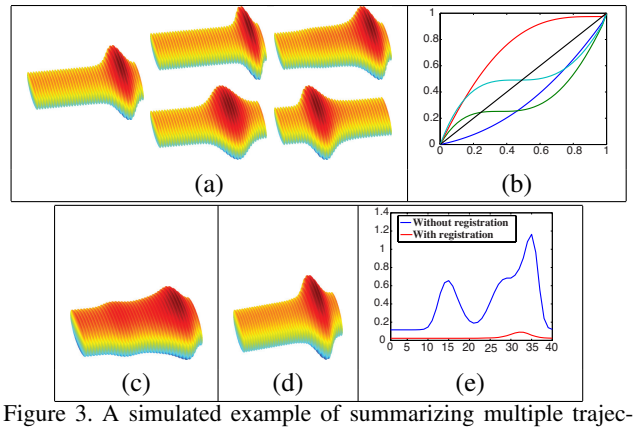


Figure 3. A simulated example of summarizing multiple trajectories. (a) and (b): simulated data and γ's; (c) and (d): mean trajectory without and with registration; (e) comparison of $\hat{\rho}$.

4. Features for audio-visual speech recognition

As we mentioned before, speech recognition is a bimodal process, involving audio and visual components. Next, we extract different features for both modalities respectively, and particularize the framework to the corresponding spaces.

4.1. Feature for audio speech recognition

The first step in any automatic speech recognition system is to extract features, i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the remaining data that carry information like background noise, emotion, etc. Mel Frequency Cepstral Coefficients (MFCCs) are commonly used as features in automatic speech and speaker recognition. For each frame, we first extract the feature of MFCCs which is a vector of coefficients. Then, the vector is scaled to be of unit norm, which makes the underlying space a unit sphere. Statistical methods for unit vectors have been studied extensively in fields of directional statistics and the landmark-based shape analysis of objects ([5]).

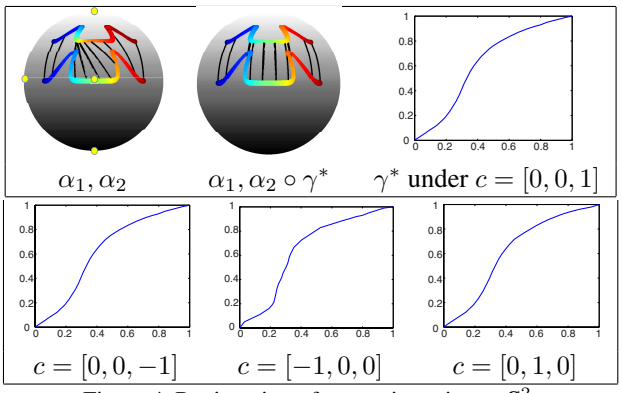


Figure 4. Registration of two trajectories on $\mathbb{S}^2$.

To apply this framework, we use the standard Euclidean Riemannian metric. For any two points $p,\ q \in \mathbb{S}^n$ $(p \neq -q)$ and a tangent vector $v \in T_p(\mathbb{S}^n)$, the parallel transport $(v)_{p\to q}$ along the shortest geodesic (i.e. great circle) from p to q is given by $v - \frac{2\langle v,q\rangle}{|p+q|^2}(p+q)$. Given two trajectories on $\mathbb{S}^n$, we use their TSRVFs and DP algorithm in Eqn. 4 to find the optimal registration between them. As an example, we show the results of registering two trajectories on $\mathbb{S}^2$ in Fig. 4. The parameterization of trajectories is displayed using colors. In the top row, the left column shows the given trajectories α_1 and α_2, the middle column shows α_1 and $\alpha_2 \circ \gamma^*$ and the right column shows γ^* using $c = [0,0,1]$. The correspondences between two trajectories are depicted by black lines connecting points along them. Due to opti-

mization of γ in Eqn. 4, the d_h value between them reduces from 1.67 to 0.36 and the correspondences become more natural after the alignment. We also try different choices of c ($c = [0, 0, -1], [-1, 0, 0], [0, 1, 0]$). The registration results are very close despite different c's as shown in the bottom row.

4.2. Feature for visual speech recognition

As we mentioned earlier, in case of the audio information is not available or listeners have incapabilities, visual information becomes important for speech recognition. Feature selection always plays an important role in image detection and classification. In this paper, we choose the covariance features for VSR. The covariance features have been widely used and proved as desired for detection and classification. Tuzel et al. [17] first introduced them for texture classification and later on extended them to tracking problems [11]. A covariance matrix of image features such as pixel location, intensity and their derivatives, is constructed to represent the image. For an image I, let $\{\mathbf{z}_k\}_{k=1\ldots n}$ be the d-dimensional feature vector at a point indexed by k in I. The image I is represented with the $d \times d$ covariance matrix of the feature points.

Several advantages of using covariance matrices as image descriptors had already been discussed in [17]. First, a single covariance matrix is typically enough to represent the image in different views and poses. Second, the covariance descriptor gives a natural way of combining multiple features which might be correlated. Third, the dimension of the covariance matrices is low compared to others. The matrix $\tilde{P}$ has only $(d^2+d)/2$ different values due to symmetry, while the representation of using raw values will need $n \times d$ dimensions.

One difficulty in analyzing covariance matrices is that the space of such matrices is not a vector space. Therefore, the traditional methods for image classification do not apply. Instead, we study these covariance matrices on a Riemannian manifold. Pennec et al. and others [9, 10] formulated this manifold as the space of nonsingular covariance matrices, i.e. symmetric positive definite (SPD) matrices. Let $\tilde{\mathcal{P}}(d)$ be the space of $d \times d$ SPD matrices and $\mathcal{P}(d) = \{P \mid P \in \tilde{\mathcal{P}}(d) \text{ and } \det(P) = 1\}$. The space $\mathcal{P}(d)$ is a well known symmetric Riemannian manifold, i.e. the quotient of the special linear group $SL(d) = \{G \in GL(d) \mid \det(G) = 1\}$ by its closed subgroup $SO(d)$ acting on the right and with an $SL(d)$-invariant metric, see [6]. Although several distances have been proposed to study elements of SPD space, very few of them come from Riemannian metrics. Furthermore, the commonly used Riemannian metric does not allow simple expression for parallel transport. So, we use the Riemannian metric introduced in [14] since the expression for parallel transport are readily available. The Lie algebra of $\mathcal{P}(d)$ is $\mathcal{T}_I(\mathcal{P}(d)) = \{A \mid A^T = A \text{ and } \text{trace}(A) = 0\}$, where I denotes the $d \times d$ identity matrix and the inner product on $\mathcal{T}_I(\mathcal{P}(d))$ is $\langle A, B\rangle = \text{trace}(AB^T)$. The tangent space at $P \in \mathcal{P}(d)$ is $\mathcal{T}_P(\mathcal{P}(d)) = \{PA \mid A \in \mathcal{T}_I(\mathcal{P}(d))\}$ and $\langle PA, PB\rangle = \text{trace}(AB^T)$. This Euclidean metric is one of the very few metrics that are invariant to the action of $SL(d)$ on $\mathcal{P}(d)$. Based on this unique metric, we have derived the required tools as follows:

- **Exponential Map**: Given $P \in \mathcal{P}(d)$ and $V \in \mathcal{T}_P(\mathcal{P}(d))$, $\exp_P(V) = \sqrt{Pe^{2P^{-1}V}P}$.
- **Inverse Exponential Map**: For any $P_1, P_2 \in \mathcal{P}(d)$, $\exp_{P_1}^{-1}(P_2) = P_1 \log(\sqrt{P_1^{-1}P_2^2P_1^{-1}})$.
- **Parallel Transport**: For any $P_1, P_2 \in \mathcal{P}(d)$, the parallel transport of $V \in \mathcal{T}_{P_1}(\mathcal{P}(d))$ from P_1 to P_2 is $P_2T_{12}^TB\ T_{12}$, where $B = P_1^{-1}V$, $T_{12} = P_{12}^{-1}P_1^{-1}P_2$ and $P_{12} = \sqrt{P_1^{-1}P_2^2P_1^{-1}}$.

In the experiments, we first compute a covariance matrix for each frame of video. To achieve enhanced robustness, especially against the diversity among different capturing environments, the covariance matrix is further normalized to a correlation matrix, where the underlying space is exactly $\mathcal{P}(d)$. A natural choice of reference point c used in Definition 1 is the identity matrix $I_{d\times d}$. For any correlation matrix $P \in \mathcal{P}(d)$, the parallel transport of $V \in \mathcal{T}_P(\mathcal{P}(d))$ from P to $I_{d\times d}$ is $P^{-1}V$ according to the equations above.

5. Experimental results

In this section, we evaluate our framework on the commonly used OuluVS dataset [19]. The OuluVS database includes 20 speakers uttering 10 phrases: *Hello, Excuse me, I am sorry, Thank you, Good bye, See you, Nice to meet you, You are welcome, How are you, Have a good time*. Each person spoke each phrase 5 times. All image sequences in the OuluVS dataset are segmented, having the mouth regions determined by the manually labeled eye positions in each frame [20]. Some examples of image sequences are shown in Fig. 5.

There are two evaluation protocols commonly adopted on the OuluVS dataset: the Speaker-Independent Test (SIT) and Speaker-Dependent Test (SDT). SIT uses the sequences from all speakers as a whole for evaluation whereas SDT separately evaluates the sequences of each speaker. In this paper, we focus on SDT to evaluate the performance. For each speaker, we compute the distance matrices and compare the rate based on rank-1 Nearest Neighbor (NN) classifier, without and with temporal alignment. The performance can be further improved using better classifiers such as SVM.

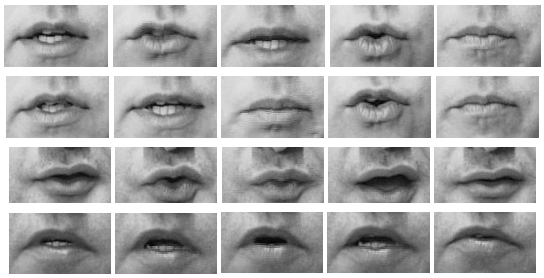


Figure 5. Example frames: First and second row: two down-sampled image sequences of the same speaker uttering the phrase "Nice to meet you"; Third and fourth row: examples of two phrases "Good bye" and "How are you", respectively by different speakers.

5.1. Performance of audio speech recognition

In this part, we extract the MFCCs from each frame, which has the dimension of 48. The underlying space becomes $\mathbb{S}^{47}$ after scaling the vectors to unit length. Then each audio can be represented as a trajectory on $\mathbb{S}^{47}$. By applying the proposed framework, we register and compute distances between trajectories of each person. Fig. 6 shows that the pattern becomes more clear after alignment for speaker 2. This is because the sequences in each class move closer after the alignment. Fig. 7 (a) shows that the classification rate increases in all of speakers after temporal alignment. The average rate increases from 55.4% to 96.0%, as shown in Table 1. The result suggests that the classification and recognition of speeches with different execution rates will benefit a lot from our method.

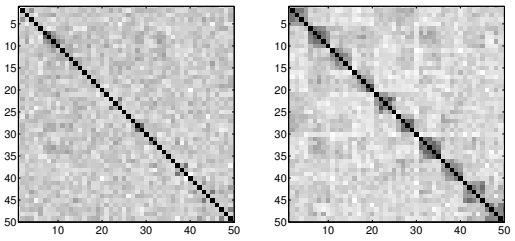
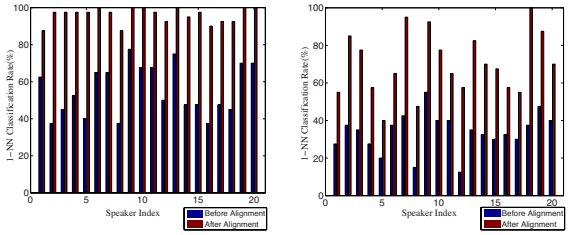


Figure 6. Distance matrix of audio features for speaker 2 (left: without alignment; right: after alignment.)

5.2. Performance of visual speech recognition

In this experiment, our definition of covariance descriptor $\tilde{P}$ for each frame is restricted to the mouth region. Seven features are extracted, including $\{x, y, I(x,y), |\frac{\partial I}{\partial x}|, |\frac{\partial I}{\partial y}|, |\frac{\partial^2 I}{\partial x}|, |\frac{\partial^2 I}{\partial y}|\}$. A 7×7 covariance matrix of these features is formed and normalized to a correlation matrix. Then for each video, we construct a trajectory of 7×7 correlation matrices. Note that one can also choose other features and obtain an improvement due to temporal alignment. We have chosen current features to demonstrate this improvement, even though these features may not be optimal in the application. We apply the framework to the space of correlation matrices and perform registration and comparison. Zhao et al. [19] had reported the SDT performance of their method on the OuluVS dataset. The rate is 70.2% based on a subset of the whole dataset, around 800 sequences by removing the short videos due to the restriction of their method. Although our method does not have such constraint, to compare in a fair way, we also remove one short video in each class for each person. Then the SDT test is performed on the remaining 800 videos. We obtain an average classification rate of 70.6% after alignment, as shown in Table 1. It is shown in Fig. 8 that the alignment has removed the temporal variabilities and distinguished different classes for speaker 18. Also, the SDT performance has been improved in all of subjects, as shown in Fig. 7 (b). Besides, we extract LBP features as introduced in [19] and compute pairwise distances using these features. The classification rate increases from 43.3% to 60.5% after alignment. In summary, in all cases of different features, the performances have drastic increases due to temporal alignment.



(a) (b)

Figure 7. Comparison of SDT performances: (a) audio features; (b) visual features.

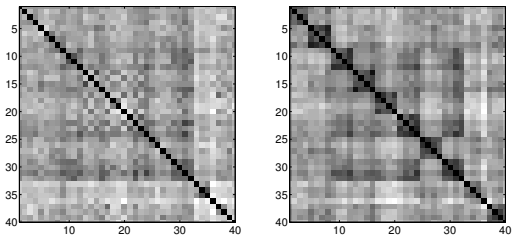


Figure 8. Distance matrix of visual features for speaker 18 (left: without alignment; right: after alignment.)

Finally, we obtain an overall performance of 97.6% by using an weighted matrix $D = wD_a + (1-w)D_v$, where D_a and D_v denotes the distance matrices obtained using audio and visual components, respectively. $w = 0.85$ is selected here. The details of classification performances are given in Table 1.

Table 1. Comparison of SDT performance on the OuluVS data.

Method		Visual	Audio	Joint
Zhao et al. [19]		70.2%	NA	NA
Our method	before alignment	33.8%	55.4%	57.6%
	after alignment	**70.6%**	**96.0%**	**97.6%**

In addition, we compute the Karcher mean and variance on this real dataset. Fig. 9 shows two examples of computing the mean trajectory: one for the seventh class of speaker 1 and the other for the first class of speaker 2. In each example, we compute the mean trajectory of five sequences and compare ρ's under two cases: without registration and with registration. It is shown that we obtain significant variance reductions after registration in both examples.

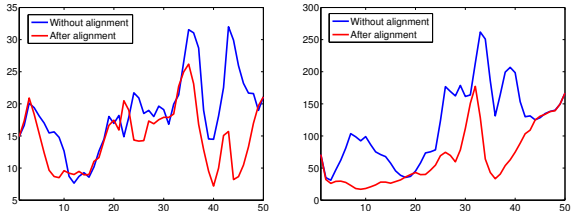


Figure 9. Comparison of ρ's in two examples. Left: the seventh class of speaker 1; Right: the first class of speaker 2.

6. Conclusion

We apply the general framework proposed in [15] to the problem of visual speech recognition. The proposed metric is proper and invariant to temporal evolutions, which allows us to register and compare trajectories simultaneously. We study the problem of speech recognition in both modalities: audio and visual components. Experimental results on the OuluVS data show that there are significant improvements of classification due to the temporal alignment. The benefit of having a proper metric also allows us to compute sample means and covariances, which could be used for classification and registration of multiple trajectories. In future work, we would like to extend the framework to other applications with different underlying manifolds, such as human activity recognition and medical imaging.

Acknowledgements

This research was supported in part by NSF grants 1217515 and 1217676.

References

[1] M. F. Beg, M. I. Miller, A. Trouve, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *IJCV*, 61:139–157, 2005. 3

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control 3rd Edition, Volume II*. Athena Scientific, 2007. 4

[3] G. E. Christensen and H. J. Johnson. Consistent image registration. *IEEE Trans. Medical Imaging*, 20(7):568–582, 2001. 3

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006. 2

[5] I. L. Dryden and K. Mardia. *Statistical Shape Analysis*. John Wiley & Son, 1998. 5

[6] J. Jost. *Riemannian geometry and geometric analysis*. Springer-Verlag, 1998. 6

[7] P. E. Jupp and J. T. Kent. Fitting smooth paths to spherical data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(1):34–46, 1987. 3

[8] H. Le and A. Kume. The Fréchet mean shape and the shape of the means. *Advances in Applied Probability*, 32(1):101–113, 2003. 4

[9] M. Liu, B. C. Vemuri, and R. Deriche. A robust variational approach for simultaneous smoothing and estimation of DTI. *NeuroImage*, 67(0):33 – 41, 2013. 6

[10] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *IJCV*, 66:41–66, 2006. 6

[11] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on Lie algebra. In *CVPR*, 2006. 6

[12] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audio-visual speech. In *IEEE*, 2003. 2

[13] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in Euclidean spaces. *IEEE Trans. PAMI*, 33:1415–1428, 2011. 3

[14] J. Su, I. Dryden, E. Klassen, H. Le, and A. Srivastava. Fitting smoothing splines to time-indexed, noisy points on nonlinear manifolds. *Image and Vision Computing*, 30(6-7):428–442, 2012. 6

[15] 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. 1, 2, 4, 8

[16] J. D. Tucker, W. Wu, and A. Srivastava. Generative models for functional data using phase and amplitude separation. *CSDA*, 61:50–66, 2013. 3

[17] O. Tuzel, F. Porikli, and P. Meer. Region covariance: a fast descriptor for detection and classification. In *ECCV*, 2006. 6

[18] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, and R. Chellappa. Rate-invariant recognition of humans and their activities. *IEEE Trans. Image Processing*, 8(6):1326–1339, 2009. 1

[19] G. Zhao, M. Barnard, and M. Pietikäinen. Lipreading with local spatiotemporal descriptors. *IEEE Trans. Multimedia*, 11(7):1254–1265, 2009. 2, 6, 7, 8

[20] G. Zhao, M. Pietikäinen, and A. Hadid. Local spatiotemporal descriptors for visual recognition of spoken phrases. In *Intl. Workshop on Human-Centered Multimedia*, 2007. 6

[21] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *CVPR*, 2012. 1