Rate-invariant recognition of humans and their activities

Ashok Veeraraghavan, *Student Member, IEEE,* Anuj Srivastava, *Member, IEEE*
Amit K. Roy-Chowdhury, *Member, IEEE* and Rama Chellappa, *Fellow, IEEE*

Abstract

Pattern recognition in video is a challenging task because of the multitude of spatio-temporal variations that occur in different videos capturing the exact same event. While traditional pattern-theoretic approaches account for the spatial changes that occur due to lighting and pose, very little has been done to address the effect of temporal rate changes in the executions of an event. In this paper, we provide a systematic model-based approach to learn the nature of such temporal variations (time warps) while simultaneously allowing for the spatial variations in the descriptors. We illustrate our approach for the problem of action recognition and provide experimental justification for the importance of accounting for rate variations in action recognition.

The model is composed of a *nominal activity trajectory* and a *function space* capturing the probability distribution of activity-specific time warping transformations. We use the square-root parameterization of time warps to derive geodesics, distance measures and probability distributions on the space of time warping functions. We then design a Bayesian algorithm which treats the execution rate function as a nuisance variable and integrates it out using Monte Carlo sampling, to generate estimates of class posteriors. This approach allows us to learn the space of time warps for each activity while simultaneously capturing other intra- and inter-class variations. Next, we discuss a special case of this approach which assumes a uniform distribution on the space of time warping functions and show how computationally efficient inference algorithms may be derived for this special case. We discuss the

Ashok Veeraraghavan is currently with Mitsubishi Electric Research Labs, Cambridge, MA 02138. Email: veerarag@merl.com. He was with Centre for Automation Research and Electrical and Computer Engineering Department, University of Maryland, College Park during the time of this work. Anuj Srivastava is with the Department of Statistics, Florida State University, 106D OSB, FSU, Tallahassee, FL. Email: anuj@stat.fsu.edu. Amit K. Roy-Chowdhury is with the Deaprtment of Electrical Engineering at the University of California, Riverside, CA. Email: amitrc@ee.ucr.edu. Rama Chellappa is with Centre for Automation Research and Electrical and Computer Engineering Department, University of Maryland, College Park, MD - 20742. Email: rama@cfar.umd.edu. This work was partially supported by an ONR MURI Grant N00014-08-1-0638. Amit K. Roy-Chowdhury was partially supported by NSF Grant IIS-0712253.

relative advantages and disadvantages of both approaches and show their efficacy using experiments on gait-based person identification and activity recognition.

I. INTRODUCTION

Pattern Recognition in videos is gaining momentum in recent years because of its applicability to several problems such as gait-based person identification, activity modeling and recognition, video-based face recognition etc. Pattern recognition in video streams is often a very challenging task because of the multitude of spatiotemporal changes that can occur in a video capturing the exact same event. Several algorithms and methods account for the spatial variations due to changes in lighting, pose and appearance of individual objects. Nevertheless, very little work has been done to account for the complex temporal variations that occur in videos. For example, in activity recognition, different instances of the same activity may consist of varying relative speeds at which the actions are executed, in addition to other intra- and inter- person variabilities. Most existing algorithms for activity recognition are not very robust to intra- and inter-personal changes of the same activity, and are sensitive to warping of the temporal axis due to variations in speed profile.

A. Prior Work in Activity Recognition:

One of the earliest investigations about the nature of human movement was the study done by photographers Etienne Jules Marey and Eadweard Muybridge [1] in the 1850s. They captured photographs of several moving subjects that revealed various interesting aspects of human and animal locomotion. The classic Moving Light Display (MLD) experiment of Johansson [2] provided a great impetus to the study and analysis of human motion perception in the field of neuroscience. This then paved the way for mathematical modeling of human action and automatic recognition, which naturally fell into the purview of computer vision.

Activity recognition has attracted tremendous interest in recent years because of its potential in applications such as surveillance, security, and human body animation. Activity recognition has been a research area since the 90's. The reader can refer to the different surveys [3][4][5] on activity recognition for a detailed review of previous research in this area. The important issues that arise in an action recognition system are discussed in detail in [3]. Broadly, action recognition has either been studied using probabilistic graphical models such as hidden Markov models [6][7][8][9] and dynamic Bayesian networks [10][11][12][13][14]. Since our approach is an attempt to account for the variabilities that affect action recognition, we provide a more

indepth coverage of prior work in this area. Recently, [15] has explicitly enumerated the three most important sources that contribute to variabilities in human activity videos as a) Viewpoint change, b) Anthropometry of actors and c) Execution rate.

1) Viewpoint and Anthropometry: Typical approaches for human action recognition begin by extracting features from a single frame or a small set of frames. These features could be simple motion-based features such as optical flow [16], and point trajectories [17], or simple silhouette-based features such as binary background subtracted images [18] or shape features [19]. Irrespective of the actual feature used for representation, it becomes important to ensure that these features are then invariant to viewpoint of the camera and the body stature of the subject (anthropometry). Most approaches use simple scaling based laws to account for anthropometry while more sophisticated approaches including affine invariance are required in order to account for view invariance. Since the focus of this paper is on modeling temporal rate variations we refer the reader to some recent methods on tackling viewpoint variations [17][20][21][22][23] and anthropometry variations [24].

2) Execution Rate: Inspite of this large body of work in accounting for viewpoint and anthropometry invariance very little has been done to account for the variability in the execution rate of the actors. Results on gait-based person identification shown in [25] indicate that it is very important to take into account the temporal variations in the person's gait. In [26], we showed that accounting for execution rate enhances recognition performance for action recognition. Typical approaches for accounting for variations in execution rate are either directly based on the dynamic time warping (DTW) algorithm [27] or some variation of this algorithm [26]. A method for computing an average shape for a set of dynamic shapes is provided in [28]. A method to learn the best class of time-warping transformations for a given classification problem is proposed in [29].

In this paper, we study the variations due to execution rate in a systematic way. We model an action sequence as a composition of these two sources of variability - variability on the feature space and variability due to execution rate. By keeping the model on the feature space completely independent of the model on the space of execution rates, we are then able to exploit any of the above mentioned viewpoint invariant features in our method. Therefore, as more sophisticated features become available our model will be able to exploit the characteristics of those features while retaining the ability to deal with variations in execution rate. We explicitly model execution rates and derive a Baysian classification algorithm for action recognition. If the chosen features are viewpoint and anthropometry invariant, then the resulting algorithm

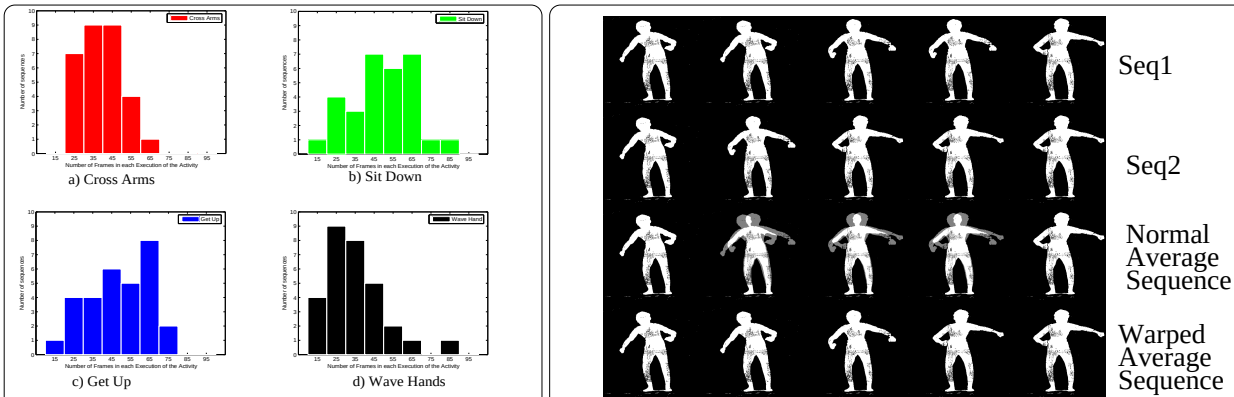
becomes invariant to all the three significant modes of variations - viewpoint, anthropometry and execution rate. Moreover, since the model developed is general and not necessarily restricted to action recognition, we believe that similar models may be used for other applications that require rate-invariance.

Motivation: Consider the INRIA iXmas activity recognition dataset. Shown in Figure 1(L) is the distribution of the number of frames in different executions of the same activity for four distinct activities. Figure 1(L) clearly shows that for the same activity the rate of execution and consequently the number of frames during the execution varies significantly. Moreover, in most realistic scenarios this temporal warping might also be inherently non-linear making simple resampling methods ineffective. This implies that for uncontrolled scenarios the variations due to temporal warpings could be even more significant. Ignoring this temporal warping might lead to structural inconsistencies apart from providing poor recognition performance. The sequence of images shown in the first two rows of Figure 1(R) correspond to two different instances of the same individual performing the same activity. There is an obvious temporal warping between the two sequences. If this temporal warping is ignored, the distance between these two sequences will be large, leading to incorrect matching. Moreover, if we are looking for some statistical description of the activity like an average sequence, ignoring the temporal warping could lead to structural inconsistencies like the presence of four arms and two heads in the average sequence as shown in the third row of Figure 1(R). If we do account for temporal warping then such inconsistencies are avoided and the distance between the two sequences is rightly small. The fourth row shows a typical average sequence obtained by our method after accounting for time warping.

Why should the distribution of time-warps be class-specific? To answer this, let us consider the activity of ‘jumping’. The subject may in principle speed up certain portions of the activity relative to the others. But, during the actual moments the subject has no contact with the ground, the only external forces on the subject are those from gravitation and therefore, much as he/she might attempt to, he/she will not be able to change the execution speed during such times. There are thus physical, aesthetic and structural constraints that force different activities to have different warping functions. The constraints themselves vary with each activity and therefore the eventual probability distribution on warping functions varies from one activity to another.

B. Contributions of the paper

- We propose a systematic generative model for activities that accounts for variations in



(L) Histogram of number of frames

(R) Two sequences with differing rates of executions and their normal and warped average sequences.

Fig. 1. (L) Histogram of the number of frames in different executions of the same action in the INRIA iXmas dataset. The histograms for 4 different activities are shown. (a) Cross Arms (b) Sit Down (c) Get Up (d) Wave hands. (R) Row 1, Row 2: Two instances of the same activity. Row 3: A simple average sequence. Row 4:Average Sequence after accounting for time warps.

speed profile of an activity. The model is composed of a *nominal activity trajectory* and a probability distribution on the *function space* of temporal warpings capturing the permissible activity-specific time warping transformations. We then derive a Bayesian solution for a rate-invariant classification of activities.

- We highlight a special case of this approach where we assume a uniform distribution on a convex subset of the warping functions and derive computationally efficient algorithms for learning and inference.
- While the preliminary conference publication [26] dealt only with uniform distribution on the space of time-warps here in this paper we deal with the learning and inference problems for a much more general class of distributions. Further, while [26] directly works with the time warping functions, we show how one can efficiently impose a Riemannian metric and perform exact and efficient statistical inference efficiently and correctly using the square-root density form of the time warp functions.

C. Outline of the paper

We begin by providing a formal statement of the problem addressed in this paper in Section II. Section III describes the geometry of the space of time warps and presents algorithms for computing geodesics, distances and prior probability distributions on this space. Section IV

describes how these tools may be used in order to learn the model parameters. Section V describes the special case of the model when the probability distribution on the space of time warpings is uniform. In Sections VI and VII, we discuss how both the models developed earlier can be used in a Bayesian recognition framework in order to perform activity analysis, recognition and activity-based person identification. Finally, in Section VIII, we present conclusions and future research directions.

II. Problem Statement

Let $C_1, C_2, ..., C_M$ be M classes (in our case M different activity labels). Here we wish to tackle two tasks while accounting for time-warping: 1. Given several instances of an activity, we would like to build a model for that activity and 2. Given a test sequence, we would like to classify the sequence to one of the models in the database.

A. Feature for representation

Observations of an activity are typically obtained using video cameras and they are in the form of video frames. Raw videos are not appropriate features for representation. In principle, the feature chosen to describe the action units must have physical significance and one must be able to directly identify the relationship between the features extracted and the basic human pose. For the problem of activity recognition, 3-D joint angles would be ideal features. Unfortunately, estimating features like 3-D joint angles from images is difficult and unreliable. So researchers have used several other features for describing the action units [17][28][30][31]. Since the USF gait database consists of monocular video, we use the shape of the silhouette (along with the appropriate Procrustes distance) as a feature [32] for the gait-based person identification experiments. The INRIA iXmas dataset contains synchronized videos from multiple views and therefore allows us to use 3D Fourier based shape features described in [33]. We refer the interested reader to [34][32] and [33] for details about the shape feature and the 3D circular FFT feature respectively.

For now let us assume that for each frame of the video, an appropriate feature has been extracted and that the video data has now been converted into a feature sequence given by $f^1, f^2, ...,$ for frames $1, 2, ...$ respectively. We will use $\mathcal{F}$ to denote the feature space associated with the chosen feature.

B. Model for warping functions

Let γ be a diffeomorphism (A diffeomorphism is a smooth, invertible function with a smooth inverse.) from $[0,1]$ to itself with $\gamma(0) = 0$ and $\gamma(1) = 1$. Also, let $\mathbf{\Gamma}$ be the set of all such functions. We will use elements of $\mathbf{\Gamma}$ to denote time warping functions. Our model for an activity consists of an average activity sequence given by $a : [0,1] \rightarrow \mathcal{F}$, a parameterized trajectory on the feature space. Any time-warped realization of this activity is then obtained using:

$$r(t) = a(\gamma(t)), \quad \gamma \in \mathbf{\Gamma} \ . \tag{1}$$

We note in passing that $\mathbf{\Gamma}$ is a group with composition as the group operation and the function $\gamma(s) = s$ as the identity element. Equation (1) actually defines an action of $\mathbf{\Gamma}$ on $\mathcal{F}^{[0,1]}$, the space of all continuous activities. In our model, the variability associated with γ in each class will be modeled using a distribution P_γ on $\mathbf{\Gamma}$. For the convenience of analysis and computation (refer Section III), we prefer to work with $\psi = +\sqrt{\dot{\gamma}}$ instead of γ directly. There is a bijection between γ and ψ and the probability models on ψ directly relate to equivalent models on γ. Thus, we will introduce probability distributions P_ψ on the set of all ψs, for each activity class.

The parameters of this model are $a(t)$, the nominal activity trajectory, and P_ψ, the probability distribution on square-root representations of time warping functions. In general, the nominal activity trajectory $a(t)$ can also be chosen to be random. But, here, we restrict our analysis to cases where, the nominal activity trajectory $a(t)$ is deterministic but unknown. We will consider parametric forms of densities for P_ψ and reduce the problem of learning P_ψ to one of learning the parameters of the distribution P_ψ. In particular, we highlight (in Section V) a special-case of a uniform distribution on the space of time warpings. This particular special-case appeared as a preliminary conference paper [26].

Physical Significance of the Model: *The nominal activity trajectory*, $a(t)$ and the probability distribution on the space of time-warps, P_ψ together capture all the possible realizations of the activity and provide the description of the activity under different variations. In general, the nominal activity trajectories of two different activities will be vastly different. The nominal activity trajectory for 'walking' would consist of key postures like heel-strike, toe-off, mid-stance etc., while that of 'sit down' would consist of the following actions - bend knee, lower body, settle on chair and rest back on backrest. The distribution of activity-specific temporal warpings P_ψ, represents the space of all permissible time-warping transformations for each activity. By learning this space, we are able to 'interpolate' appropriately between training sequences. Suppose there is a test sequence that is within this space, but was not a part of the training sequences. Most

template sequence-based recognition techniques tend to misclassify such test sequences. Learning the function space of an activity provides our algorithm with the generalization power necessary to correctly classify such test sequences. Moreover, by formally learning this warping space in a class specific manner, we also obtain better discriminative power than other heuristic techniques for handling time-warping. The model M=$\{a, P_\psi\}$ represents a *function space* of activities whose elements are composed of functions $a(\gamma(t)),\ \forall\ \gamma \in \mathbf{\Gamma}$.

C. Problems

Here, we state informal descriptions of the various problems we wish to tackle.

1) The Learning Problem: Given N labeled realizations $r_1, r_2, r_3, ...r_N$, of an activity, we would like to learn the model for this activity. This is equivalent to learning the nominal activity trajectory $a(t)$ and the distribution on the warping parameters given by P_ψ.

2) The Classification Problem: Suppose we have models for M different activities $\{a^i, P^i_\psi\}_{i=1}^M$. Given a test sequence $r(t)$, we would like to classify this test sequence as belonging to one of the M models.

3) Clustering Problem: Given several realizations from K different activities with no class labeling, we would like to cluster these sequences into K distinct clusters such that sequences within the same cluster are maximally similar while sequences in different clusters are dissimilar. Moreover, unlike traditional clustering algorithms this similarity is invariant to changes in exectution rate of the action since the model for each cluster is built to be rate-invariant.

III. Differential geometric tools on the space of time-warping functions

The model for a random observation of an activity class consists of $a(\gamma(t))$, where a is the average of that class and γ is a warping function. In order to classify activities at variable execution rates, we need to analyze the warping functions as random functions. However, the space of warping functions is not a vector space and that rules out the use of classical functional analysis for this task. One alternative is to utilize the differential geometry of this space, impose a Riemannian structure on it, and use appropriate tools to perform calculus and statistics of warping functions. In particular, we can compute distances between warping functions, estimate sample means for given warping functions, and impose parametric and non-parametric probability distributions on the space of warping functions.

The next question is: What Riemannian structure on the space of warping functions is suitable and convenient for activity recognition? The Fisher-Rao metric is often used for analyzing

probability density functions. (The Cramer-Rao lower bound on estimation of parameters is derived using this metric.) One major reason for its popularity is that it is invariant to arbitrary warpings of the functions involved. In other words, under this metric the distance between any two warping functions $\gamma_1(t)$ and $\gamma_2(t)$ is same as that between $\gamma_1(\gamma(t))$ and $\gamma_2(\gamma(t))$ for any arbitrary warping function $\gamma(t)$. This point is important in activity recognition because, as we will point out in Section IV-D, the representation of an activity model is not unique, i.e. there is no canonical choice of γ for representing activity models. The choice of Fisher-Rao metric implies that the resulting distances are same irrespective of the baseline time axis chosen to represent activity models.

The Fisher-Rao metric, when applied to different mathematical representations of γ, i.e. γ, $\dot{\gamma}$, $\log \dot{\gamma}$, or $\sqrt{\dot{\gamma}}$, takes different forms. Interestingly, in the case of $\psi \equiv \sqrt{\dot{\gamma}}$, this metric simplifies to the familiar and convenient $\mathbb{L}^2$ metric [35], [36]. Furthermore, the space of all warping functions, represented by their square-root density forms, under the Fisher-Rao metric, becomes a unit sphere. This is because

$$\|\psi\|^2 = \int_0^1 |\psi(t)|^2 dt = \int_0^1 |\dot{\gamma}(t)| dt = \gamma(1) - \gamma(0) = 1 \ .$$

For these two reasons – invariance to arbitrary time scalings and the spherical nature of the resulting space, we choose the square-root density form to represent and analyze variability associated with the warping functions.

Let the space of all square-root density forms be given by

$$\mathbf{\Psi} = \{\psi : [0,1] \to \mathbb{R} | \psi \geq 0, \int_0^1 \psi^2(t) dt = 1\} \ . \tag{2}$$

This is the positive orthant of a unit hypersphere in the Hilbert space of all square-integrable functions on $[0,1]$. Let $T_\psi(\mathbf{\Psi})$ be the tangent space to $\mathbf{\Psi}$ at any given point ψ. Then, for any v_1 and v_2 in $T_\psi(\mathbf{\Psi})$, the Fisher-Rao metric is given by

$$\langle v_1, v_2 \rangle = \int_0^1 v_1(t) v_2(t) dt. \tag{3}$$

Since $\mathbf{\Psi}$ is a sphere, its geometry is well known and we can directly use known expressions for tools such as geodesics, exponential maps, and inverse exponential maps on $\mathbf{\Psi}$. Consequently, the algorithms for computing sample statistics, defining probability density functions, and generating inferences also become straightforward.

We begin by describing some elements of differential geometry of $\mathbf{\Psi}$.

A. Geometry of $\mathbf{\Psi}$

One way to quantify the differences between two warping functions is to compute the distance between their corresponding representations in $\mathbf{\Psi}$. This distance is given by the length of a geodesic, the shortest path connecting those two points in $\mathbf{\Psi}$. We know that the geodesics on a sphere are the great circles and the geodesic distance is simply the length of the shorter arc connecting the two points on a great circle. Given two warping functions γ_1 and γ_2, and their square-root density forms, ψ_1 and ψ_2 in $\mathbf{\Psi}$, the geodesic distance between them on $\mathbf{\Psi}$ is given by

$$d(\psi_1, \psi_2) = \cos^{-1}(\langle \psi_1, \psi_2 \rangle), \tag{4}$$

where $\langle \psi_1, \psi_2 \rangle = \int_0^1 \psi_1(t)\psi_2(t)dt$.

The geodesic path itself can also be computed rather simply. Taking the radial projection of the chord joining points ψ_1 and ψ_2 onto the unit sphere results in the geodesic. The chord joining ψ_1 and ψ_2 is given by $(1-s)\psi_1 + s\psi_2$ where s is the parameter that identifies various points on this chord. The radial distance of a point on this chord is given by $s^2 + (1-s)^2 + 2s(1-s)(\langle \psi_1, \psi_2 \rangle)$. Therefore, we can analytically write the geodesic connecting ψ_1 and ψ_2 as: $X : [0,1] \rightarrow \mathbf{\Psi}$,

$$X(s) = \frac{(1-s)\psi_1 + s\psi_2}{s^2 + (1-s)^2 + 2s(1-s)(\langle \psi_1, \psi_2 \rangle)},$$

such that $X(0) = \psi_1$ and $X(1) = \psi_2$. Another way to specify a geodesic path in $\mathbf{\Psi}$ is by giving a starting point $\psi \in \mathbf{\Psi}$ and a starting direction $v \in T_\psi(\mathbf{\Psi})$:

$$X(s) = \cos(s\|v\|)\psi + \sin(s\|v\|)\frac{v}{\|v\|} \ , \tag{5}$$

where $\|v\| = \sqrt{\int_0^1 v(t)^2 dt}$.

One use of geodesics is to define and compute the exponential map from $T_{\psi_1}(\psi)$ to ψ. It is simply the value reached at $s = 1$ by a geodesic that starts from ψ in the direction v and moves at a constant speed. We can evaluate the exponential map using:

$$exp_\psi(v) = \cos(\|v\|)\psi + \sin(\|v\|)\frac{v}{\|v\|}. \tag{6}$$

Similarly the inverse of the exponential map $exp_{\psi_1}^{-1}(\psi_2) = v \in T_{\psi_1}(\psi)$ can also be computed analytically using

$$u = \psi_2 - \langle \psi_2, \psi_1 \rangle \psi_1 \tag{7}$$

$$v = \frac{u \cos^{-1}(\langle \psi_1, \psi_2 \rangle)}{\sqrt{\langle u, u \rangle}}. \tag{8}$$

B. Statistical Analysis on $\mathbf{\Psi}$

With the geometry of $\mathbf{\Psi}$ as specified above, let us derive some tools for statistical analysis of data. Given a number of observed warping functions, we will estimate the sample mean and covariance, use these estimates to define a "wrapped-Gaussian" density function and derive Bayesian classification algorithms using these densities as priors.

To compute the sample means of elements of $\mathbf{\Psi}$, we will use the notion of Karcher mean [37] that has been used frequently for defining means on nonlinear manifolds. Suppose, we have n different square-root density forms, given by $\psi_1, \psi_2, ...\psi_n$. Then, their Karcher mean $\bar{\psi}$ is defined as the element that minimizes the sum of squares of geodesic distances:

$$\bar{\psi} = \arg\min_{\psi \in \mathbf{\Psi}} \sum_{i=1}^{n} d(\psi, \psi_i)^2 \tag{9}$$

where, d is the geodesic distance defined in (4). Note that the Karcher mean may not be unique and can instead be a set of elements. A commonly used approach for finding a Karcher mean is to use the gradients and this is where the exponential map and its inverse are needed. The iterative update to the current value of mean is given by:

$$\bar{\psi} \rightarrow exp_{\bar{\psi}}(\epsilon v), \quad \text{where} \quad v = \frac{1}{n} \sum_{i=1}^{n} exp_{\bar{\psi}}^{-1}(\psi_i) \tag{10}$$

and where ϵ is usually 0.5.

The next step is to define and compute a sample covariance for the observed ψs. The key idea here is to use the fact that the tangent space $T_{\bar{\psi}}(\mathbf{\Psi})$ is a vector space. Using a finite-dimensional approximation, say $V \subset T_{\bar{\psi}}(\mathbf{\Psi})$, we can use the classical multivariate calculus for this purpose. In practice, we obtain a natural restriction when v is observed at a finite number, say T, of times leading to an observation $\{v(t_i)|i = 1, 2, \ldots, T\}$. With a slight abuse of notation, we will denote this vector by $v \in \mathbb{R}^T$. The resulting sample covariance matrix is given by: $\bar{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} v_i v_i^T$, where each v_i is a T-dimensional sample of the function $exp_{\bar{\psi}}^{-1}\psi_i$. Note that by definition, the mean of v_is should be zero. In cases where the number of samples n is smaller than T, one can apply an additional dimension-reduction tool to work on a smaller space. For instance, we can use the singular value decomposition (SVD) of the sample covariance matrix $\bar{\Sigma}$ and retain only the top m significant singular values and the corresponding singular vectors. In such cases, the covariance matrix is indirectly stored using $\lambda_1, \lambda_2, ...\lambda_m$ singular values and their corresponding singular vectors $u_1, u_2, ...u_m$.

Next, we define a "wrapped-Gaussian" probability density on $\mathbf{\Psi}$. We say "wrapped-Gaussian" because $\mathbf{\Psi}$ is a non-Euclidean space and it is not possible to define a Gaussian density here. We

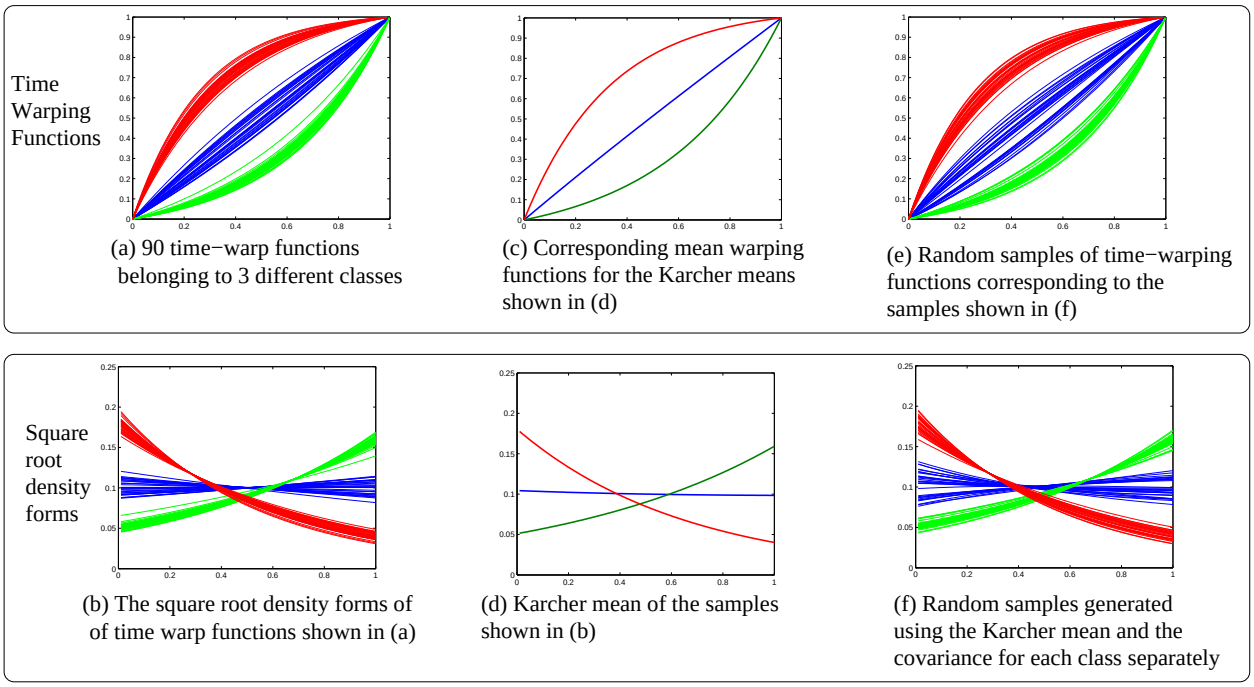
follow the tangent PCA (TPCA) approach [38] for defining probability densities on nonlinear manifolds. In this approach, one defines a Gaussian probability density on a tangent space of the manifold and then projects it onto the manifold using the exponential map. However, in our case we need only the samples from the eventual density function and the explicit functional form of that projected density is not needed. In fact, we will apply one more transformation in taking the samples on $\mathbf{\Psi}$ to obtain samples on $\mathbf{\Gamma}$. For a mean μ and covariance Σ, we can define a normal density function $N(v|\mu, \Sigma)$ on the elements of $V \subset T_\mu(\mathbf{\Psi})$. In case the data is available in the form of prior samples, we can use the sample means and covariances to define this density on the space V. The exponential map: $\exp_{\bar{\psi}} : T_{\bar{\psi}}(\mathbf{\Psi}) \rightarrow \mathbf{\Psi}$ maps this density to the spherical space of square-root forms, and the mapping $\psi \mapsto \gamma(t) = \int_0^t |\psi(\tau)|^2 d\tau$ takes it further to the space of warping functions. The exponential map results in wrapping the Gaussian density on the tangent space onto the sphere and therefore the name *wrapped-Gaussian*. We will denote the resulting densities on $\mathbf{\Psi}$ and Γ by P_ψ and P_γ, respectively.

For a Bayesian classification of activities, as described later in this paper, we will need to estimate the posterior probability of different classes given the observed data. In this calculation, the warping function is considered a nuisance variable that needs to be integrated out. Using a Monte Carlo approach, we will generate samples from the prior on γ and use those samples to approximate the nuisance integral. Thus, we have a need to generate samples from the class-specific priors P_γ on Γ. This, in turn, requires sampling from the probability density P_ψ, which is accomplished as follows. Let $\bar{\psi}$ and $\bar{\Sigma}$ be the sample mean and the sample covariance of the square-root forms observed in a particular class. Assume that the covariance is stored in the form of m singular values λ_is and corresponding singular vectors u_is. In such cases, a random sample from the model P_ψ is given as

$$\psi \sim exp_{\bar{\psi}}(v) \qquad \text{where} \quad v \sim \sum_{i=1}^{m} z_i \sqrt{\lambda_i} u_i \quad \text{and} \quad z_i \sim N(0,1) \tag{11}$$

This random sample can then be converted into a warping function using the partial integration $\psi \mapsto \gamma$ such that $\gamma(t) = \int_0^t |\psi(\tau)|^2 d\tau$.

Example Consider the example shown in Figure 2. Figure 2(a) shows 30 sample time-warping functions from each of three different classes (color coded). The corresponding square-root density forms are shown in 2(b) and can be computed using $\psi = \sqrt{\dot{\gamma}}$. For each class using the samples of the square-root density forms we can compute the Karcher mean and the covariance as in Equation 9. The Karcher means are shown in 2(d). The mean time-warping functions for each class obtained by partially integrating the Karcher means are shown in 2(c). The model



(a) 90 time−warp functions belonging to 3 different classes

(c) Corresponding mean warping functions for the Karcher means shown in (d)

(e) Random samples of time−warping functions corresponding to the samples shown in (f)

(b) The square root density forms of of time warp functions shown in (a)

(d) Karcher mean of the samples shown in (b)

(f) Random samples generated using the Karcher mean and the covariance for each class separately

Fig. 2. Figure is Color coded - Each color represents a different class (a) Random samples of time-warping functions belonging to 3 different classes (color coded) (b) Corresponding samples of square-root density forms (c) Mean time-warping function for each class computed by partial integration of the class-specific Karcher mean (d) Class specific Karcher mean computed using the samples shown in (b) (e) Random samples generated from the stored model (f) Random samples of ψ generated from the stored Karcher means and covariance.

for each class of time-warping functions is encoded in the form of the corresponding Karcher means and covariances. Now one can generate random samples from this model as described above. Shown in 2(f) are sample square-root density forms generated using the model parameters for each class (i.e., the Karcher mean and covariances). As before the corresponding time-warping functions maybe computed via partial integration and are shown in 2(e). We encourage the interested reader to download sample code either from the supplemental material or from http://www.cfar.umd.edu/~vashok/Documents/TIP_Code_Supplemental.zip to generate some of the figures and results shown in this example.

C. Global Speed of activity

We have restricted our attention to time-warping functions from $[0, 1]$ to itself, i.e the functions that do not contract or dilate the full duration of the activity. We claim that this is not restrictive, since any other time-warping transformation can be decomposed into two parts: a global linear

scaling of the temporal axis and the non-linear time-warping functions that we have addressed so far. The effect of such a linear global temporal scaling is identical to the effect of changing the rate of sampling.

Let $a(t)$, for $0 \leq t \leq T_a$, be a vector valued function of time. Let $b(t)$, for $0 \leq t \leq T_b$, be a time-warped version of $a(t)$, with the warping function given by $w(t)$, i.e., $b(t) = a(w(t))$, $w(t) : [0, T_b] \rightarrow [0, T_a]$. Now $w(t)$ can be decomposed as $w(t) = T_a\gamma(t/T_b)$ where $\gamma : [0, 1] \rightarrow [0, 1]$ i.e., a global linear dilation (or contraction) and a non-linear warping γ. Without loss of generality we will use the word time-warping transformation to synonymously denote the non-linear time warping function given by γ. In all our experiments we have first identified the global temporal scaling factor by identifying the start and stop instants of each activity. The identification of the start and stop instants of each activity is also done automatically by template matching. Once the global temporal scaling factor is found, each realization of the activity is temporally dilated or contracted linearly so that the total duration of the activity is a constant for all realizations of the activity.

IV. Learning and Classification Algorithms

Given N realizations $r_1, r_2, r_3, ...r_N$, of an activity, we need to learn the parameters of the model for this activity. This amounts to learning the nominal activity trajectory $a(t)$ and the probability distribution P_ψ.

A. Estimating P_ψ given $a(t)$

Let us assume that the nominal activity trajectory $a(t)$ is known. Now we need to estimate the parameters of the warping distribution which is given by P_ψ. In order to learn P_ψ, we first warp each of the observed realizations of the activity to the known nominal activity trajectory given by $a(t)$. This warping can be performed using the DTW algorithm. The DTW algorithm provides us with corresponding warping functions $\gamma_i(t)$ such that $\int_0^1 \|r_i(t) - a(\gamma_i(t))\|^2 dt$ is minimized. Then, we can compute ψ_is using $\psi_i = \sqrt{\dot{\gamma}_i}$.

Now, we have several samples $\psi_1, \psi_2, ...$ to estimate the distribution P_ψ. Assuming a "wrapped-Gaussian" distribution on $\mathbf{\Psi}$, this amounts to estimating the sample mean and the sample covariance of the observed ψ_is. As described in Section III-B, we can define and compute the Karcher mean of given ψ_is using the exponential and the inverse exponential maps. The covariance is obtained similarly by restricting to a T-dimensional approximation V of the vector

space $T_{\bar{\psi}}(\mathbf{\Psi})$. Using SVD of observations in V, one ends up with the singular values $\lambda_1, \lambda_2, ...\lambda_m$ and their corresponding singular vectors $u_1, u_2, ...u_m$.

Thus, given the nominal activity trajectory $a(t)$, we can estimate the parameters of the warping distribution P_ψ, namely its Karcher mean $\bar{\psi}_K$ and its covariance stored indirectly using m singular values $\lambda_1, \lambda_2, ...\lambda_m$ and corresponding singular vectors $u_1, u_2, ...u_m$.

B. Estimating $a(t)$ assuming known warping functions

For the given observations $r_1, r_2, \ldots,$ of an activity, assume that the corresponding warping functions $\gamma_1, \gamma_2, \ldots,$ are also given. Then, we can estimate the nominal or average activity trajectory $a(t)$ using

$$\bar{a}(t) = \frac{1}{N} \sum_{i=1}^{N} r_i(\gamma_i^{-1}(t)) \tag{12}$$

C. Iteratively estimating $a(t)$ and P_ψ

Given N realizations $r_1, r_2, r_3, ...r_N$, of the same activity, we would like to learn the parameters of the model for this activity. We do this by iteratively estimating P_ψ and refining our estimate of the nominal activity trajectory $\bar{a}(t)$ using the steps described in the previous two sections. We first initialize the nominal activity trajectory to one of the realizations say $a_{init}(t) = r_1(t)$. Then we estimate P_ψ using the method described in Section IV-A. We then refine the estimate of the nominal activity trajectory using the method described in Section IV-B. These two steps are iterated till convergence. In practice, we find that the iterations converge very quickly (within 4 or 5 iterations).

D. Uniqueness of the Model parameters

The model parameters given by $a(t)$ and $P_\psi \approx \{\bar{\psi_K}, \Sigma_\psi\}$ are not unique. Two different sets of model parameters $M_1 = \{a_1(t), P_{\psi_1}\}$ and $M_2 = \{a_2(t), P_{\psi_2}\}$, could lead to the same distribution on the observation space. That is, the two models may lead to the same distribution on the space of all activity realizations. This could happen if the corresponding nominal activity trajectory and the distribution on the space of warping transformations are related as

$$a_2(t) = a_1(\gamma(t)) \qquad \bar{\psi_1} = \sqrt{\dot{\bar{\gamma}_1}} \qquad \bar{\psi_2} = \sqrt{\dot{\bar{\gamma}_2}} \qquad \bar{\gamma_2}(t) = \bar{\gamma_1}(\bar{\gamma}^{-1}(t)) \qquad \Sigma_2 = \Sigma_1 \tag{13}$$

When the conditions listed in (13) are satisfied, we notice that $a_2(\bar{\gamma}_2(t)) = a_1(\bar{\gamma}_1(t))$, i.e., the mode of the activity trajectories is the same for both models. Moreover, since the covariance

matrices for the two models are identical ($\Sigma_1 = \Sigma_2$), this means that samples for either of these models will have identical distributions and would therefore be indistinguishable. In practice this means that there is an equivalence class of models such that any two models from the same equivalence class are indistinguishable. The conditions for belonging to the same equivalence class are those stated in (13). While performing classification and inference based on these model parameters it becomes essential to maintain uniqueness of model parameters. Therefore, once we learn the model parameters we always choose a single canonical representation for each equivalence class. Note that the choice of this canonical representation does not affect the performance of the algorithm at all as long as this choice is consistent. We choose the model with $\bar{\gamma}_K(t) = t$, such that the Karcher mean of the warping distribution corresponds to simple linear warping and the covariance matrix of the warping transformations encodes all the non-linearities in the warping distributions. The canonical model parameters are unique and can be directly used for classification and inference.

E. Generating activity samples from the model

The model for an activity is given by the nominal activity trajectory $a(t)$ and the distribution on warping transformations given by P_ψ. We can use this model to generate random samples from the model. We first generate random samples $\psi_1, \psi_2, ...\psi_M$ from the warping distribution P_ψ as described in Section III-B. The corresponding time warp for each ψ is computed. Let $\gamma_1, \gamma_2, ...\gamma_M$ be the corresponding time warps. Then realizations from the model may be drawn as

$$r_j(t) = a(\gamma_j(t)) + w(t) \qquad \text{where} \qquad w \backsim N(0, \Sigma). \tag{14}$$

F. Classification Algorithm

Let us assume that we have K different models $M_1, M_2, ...M_K$ given by their appropriate nominal activity trajectories $a_1, a_2, ...a_K$ and corresponding P_ψ given by $P^1_\psi, P^2_\psi, ..., P^K_\psi$. Given a test sequence $r(t)$, we would like to classify $r(t)$ to one of the K possible classes. This classification task can be accomplished using MAP estimation, i.e.,

$$ID = \arg \max_{i=1,2,..K} P(M_i|r) = \arg_{i=1,2,..K} \max P(r|M_i)P(M_i). \tag{15}$$

The likelihood $P(r|M_i)$ can be computed as,

$$P(r|M_i) = \int_\psi P(r|M_i, \psi)P(\psi|M_i)d\psi \qquad \text{where} \qquad P(\psi|M_i) = P^i_\psi. \tag{16}$$

This integral can be estimated using Monte Carlo sampling methods. We draw N samples from the model M_i as described in Section IV-E. Using these samples we estimate the likelihood $P(r|M_i)$ as

$$P(r|M_i) = \frac{1}{N}\sum_{j=1}^{j=N} P(r|a_i, \psi_j) \qquad \text{where} \qquad \psi_j \backsim P(\psi) = P_\psi^i \tag{17}$$

In order to compute the summation described above, we need a model for computing the conditional likelihood $P(r|M_i, \psi_j)$. The conditional warp probability is inversely proportional to the squared distance between the warped nominal activity trajectory and the test sequence, i.e.,

$$P(r|M_i, \psi_j) = e^{-\alpha D(r, a_i(\gamma_j))} \qquad \text{where } D(r, a_i(\gamma_j)) = \int_0^1 (r(t) - a_i(\gamma_j(t)))^2 dt \tag{18}$$

and α is a suitably chosen constant. As the number of samples N increases the accuracy of the approximation improves. One can also improve the accuracy of the approximation by performing importance sampling [39]. Let us assume that the proposal distribution from which the samples of the ψ are drawn is given by $G(\psi)$. Then we draw N samples of ψ from G and the integral is approximated as

$$P(r|M_i) = \frac{1}{N}\sum_{j=1}^{j=N} P(r|M_i, \psi_j)\frac{P(\psi_j|M_i)}{G(\psi_j)} \qquad \text{where} \qquad \psi_j \backsim G(\psi). \tag{19}$$

In practice, using importance sampling significantly improves the accuracy of the approximation when using a finite number of samples. The effectiveness of importance sampling also critically depends upon the proposal distribution. The proposal distribution or the importance distribution should ideally be as close to the posterior distribution we wish to approximate. In practice, we first estimate the mode of this posterior by computing the best warping transformation between the nominal activity trajectory of the model ($a_i(t)$) and the test sequence ($r(t)$). We set the mean of the importance distribution to be this warping transformation while letting the covariance of the importance distribution to be the same as the covariance of the model. We have experimentally found that this choice of importance distribution enables us to effectively approximate the integrals using Monte Carlo methods with a reasonable number of random samples.

V. Function Space of Time-Warps

The model described in the previous sections represents an activity using a nominal activity trajectory $a(t)$ and a probability distribution on the space of time warpings P_ψ. There are two inherent difficulties in practical implementations of such a model inspite of its rigour. Firstly,

since the model attemps to learn a probability distribution on the space of permissible time-warping functions, the algorithm for learning this P_ψ requires a reasonable number of sample realizations of each action. In the presence of very few samples, the learning algorithm might lead to underfitting of the data. Moreover, as inference using this model is done using Monte Carlo methods, the algorithms for inference are computationally expensive.

Suppose we relax the assumption about learning the probability distribution of permissible time-warps and instead attempt to learn a subset in the time-warping space and assume that the probability distribution of time-warps is uniform within the learnt subset. Each activity can now be represented by using a nominal activity trajectory given by $a(t)$ and W, the set containing all the time warping transformations permissible for that activity. Each realization of an activity is given by a trajectory $r(t) = a(f(t))$ where $f \in W$. Such a model is a special case of learning P_ψ where, we assume that the probability distribution is uniform on a subset $W \in \Gamma$ in the space of time-warpings. The advantage of using such a model where the probability distribution is assumed uniform is that both the learning and the inference algorithms become simple dynamic programming problems when we constrain the set W to be a convex set.

A. Activity specific time-warping space (W)

Even though $\mathbf{\Gamma}$ represents the space of all plausible time-warping transformations, every individual activity may only be able to access a subset W of the candidate functions in $\mathbf{\Gamma}$ because of the physical constraints imposed on the actor and the activity. We can then model the activity using a uniform distribution on this subset W. Then learning the parameters of the uniform distribution boils down to learning this subset W. Below, we discuss and visualize some properties of this activity specific time warping space W.

1) W is a subset of $\mathbf{\Gamma}$, i.e., $W \subset \mathbf{\Gamma}$.
2) $\gamma(t) = t$ is a candidate function in W, i.e., $\gamma(t) = t \in W$. This represents no time warping.
3) It is reasonable to assume that W is convex, i.e., $\forall\ \gamma_1, \gamma_2 \in W$ and $\alpha \in (0, 1)$, $\gamma = \alpha\gamma_1 + (1-\alpha)\gamma_2 \in W$. Since the derivative is a linear operator, this means that if the rate of execution of some action unit can be speeded up by factors α_1 and α_2 then it can also be speeded up by any factor β in between α_1 and α_2. This is not just reasonable but in fact desirable.

This implies that W can be bounded above and below by functions u, $l \in W$ such that

$$u(t) \geq t \geq l(t) \ \ \forall t \in (0,1) \qquad \text{and} \qquad u \geq \gamma \geq l \ \ \forall \gamma \in W \tag{20}$$

where $\gamma_1 \geq \gamma_2 \Longrightarrow \gamma_1(t) \geq \gamma_2(t) \ \ \forall t \in (0,1)$. So, we can now index any such convex space W by the functions u and l and call it W_{ul} and learning W is essentially the same as learning the

upper and lower bounding functions u and l.

B. Symmetric representation of an Activity Model

As described for the "wrapped-Gaussian" distribution, the representation of the activity model given by $M_1 = \{a(t), W_{ul}\}$ is not unique. Let $u_{new}(t) = f^{-1}(u(t))$ and $l_{new}(t) = f^{-1}(l(t))$ and let f be a member function in W_{ul}. Consider the new model $M_2 = \{b(t), W_{u_{new}l_{new}}\} = \{a(f(t)), W_{u_{new}l_{new}}\}$. For every realization of the model M_1, i.e., $a(\gamma_1(t))$ there exists a corresponding realization of the model M_2 given by $b(f^{-1}(\gamma_1(t)))$. Therefore the two models M_1 and M_2 are equivalent. As before, we will resolve this ambiguity by specifying a synchronizing time such that the average of all the warping functions in W_s is the identity warping function. The *symmetric* representation of the model is such that $u_{new}(t) - t = t - l_{new}(t)$. Therefore the activity specific warping space can be represented as $W_s = W_{u_{new}l_{new}}$ where $s(t) = u_{new}(t) - t = t - l_{new}(t)$, represents the extent of possible temporal warpings. This symmetric representation of the model is unique, i.e., if $M_1 = \{a_1(t), W_{s1}\}$ and $M_2 = \{a_2(t), W_{s2}\}$, then $M_1 = M_2 \iff a_1 = a_2$ and $s_1 = s_2$.

Given a non-symmetric representation of the model, i.e., $M_1 = \{a(t), W_{ul}\}$, we still need to determine a time-warping function f such that upper and lower bounding functions of the new model are symmetric about the diagonal. This is achieved as

$$\begin{aligned} u_{new}(t) - t \quad & = t - l_{new}(t) \qquad (21) \\ \text{(Substituting for } u_{new}(t) \quad & \text{and applying the } u^{-1} \text{ operator)} \\ \Rightarrow \qquad f(t) = \quad & \{2u^{-1}(t) - f^{-1}(l(u^{-1}(t)))\}^{-1} \end{aligned}$$

This implicit function equation can be solved by fixed point iterations as $f_{(i)}(t) = \{2u^{-1}(t) - f^{-1}_{(i-1)}(l(u^{-1}(t)))\}^{-1}$, where $f_{(i)}$ represents the approximation of f in the i^{th} iteration. We initialize the iteration with $f_{(0)}(t) = \frac{u(t)+l(t)}{2}$. We observe that it converges within very few iterations with such an initialization. Once we have obtained this symmetrizing time warp f then any non-symmetric model parameters $M_1 = \{a(t), W_{ul}\}$ can be transformed to its symmetric (unique) counterpart as $M = \{b(t), W_s\}$, where $b(t) = a(f(t))$ and $s(t) = u_{new}(t) - t = t - l_{new}(t) = f^{-1}(u(t)) - t$.

C. Learning Model Parameters

Learning the model parameters can be done as before by iterating between the two unknowns ($a(t)$ and P_γ). Learning the nominal activity trajectory $a(t)$ is done as described in Section IV-B. The only difference between earlier and now is during the estimation of the parameters P_ψ. Earlier we computed the Karcher mean and the covariance of P_ψ for the wrapped-Gaussian distribution, here since the parameters of P_γ are given by the upper and lower bounding functions we need to estimate them. Given an estimate of the activity trajectory $a(t)$ and corresponding warping functions $\gamma_i(t)$ for each realization, the the upper and the lower bounding functions for the activity specific time-warping set can be estimated as

$$\hat{u}(t) = \max_{i=1,2,...N} \gamma_i(t),\ \forall t \in (0,1) \qquad \text{and} \qquad \hat{l}(t) = \min_{i=1,2,...N} \gamma_i(t),\ \forall t \in (0,1). \tag{22}$$

Since each γ_i is constrained to be monotonously increasing and the end points are fixed, it is easy to see that the estimates $\hat{u}(t)$ and $\hat{l}(t)$ also inherit these properties. Thus the estimated model $\hat{M}$ is given by $\hat{M} = \{\hat{b}(t), W_{ul}\}$. This model parameters correspond to the non-symmetric version of the model and can be easily transformed to the equivalent symmetric version of the model using the procedure described in Section V-B.

D. Classification using the model

The primary advantage of using the uniform distribution on the space of time-warping functions instead of learning a class-specific probability density function is that the classification algorithm becomes computationally efficient. While classification in the general case is dependent on Monte-carlo methods, we show how a simple dynamic programming based algorithm will suffice for classification using the uniform distribution based model. Suppose we have M different activity models given by $M_i = \{a_i(t), W_{s_i}\}$ for $i = 1, ..M$. Given a test sequence $h(t)$, the activity recognition problem is one of identifying the model that generated the test sequence $h(t)$. We do this in two steps. Firstly, assuming that the test sequence $h(t)$ is generated from the model M_i, we estimate the best warping transformation $\hat{f}_i$ from W_{s_i} that would warp a_i to h, i.e.,

$$\hat{f}_i = \min_{f \in W_{s_i}} dist(h(t), a_i(f(t))) \tag{23}$$

$$\hat{I} = \arg\min_{i=1,...M} dist(h(t), a_i(\hat{f}_i(t))) \tag{24}$$

Activity recognition is performed by minimizing the warping error between the nominal activity trajectory and the test sequence. Note that the search of warping functions is performed only over the corresponding activity specific warping set. The above-mentioned intuitive idea for

activity recognition can be easily implemented by a simple variation of the DTW. In the DTW algorithm, instead of arbitrarily limiting the warping function to lie within some window (typical choices are uniform window and parallelogram window), we replace the window constraints by the upper and lower bounds for the warping function that we have learnt for each model. Thus, the DTW algorithm with the window width being given by $u(t) = s(t) + t$ and $l(t) = t - s(t)$ computes the distance that is being minimized in (24).

$$\hat{I} = \min_{i=1,\ldots M} DTW(a_i, h, s) \tag{25}$$

where, $DTW(a_i, h, s)$ stands for the implementation of the DTW algorithm with the warping window constraints given by $u(t) = s(t) + t$ and $l(t) = t - s(t)$.

VI. Experiments

We tested the algorithms on three different datasets - UMD Common Activities dataset, the INRIA iXmas dataset and the USF gait dataset. We used a warpped-Gaussian probability distribution for P_ψ with its parameters stored using a set of tangent plane vectors u_ψ and their covariance matrix Σ_ψ. We denote the experimental results using this algorithm as P_{Gauss} in the results. We also implemented the uniform distribution on the space of time-warping functions using dynamic programming and performed maximum likelihood inference using this model. We denote the results using this method as P_{Unif} in the results.

A. Common Activities Dataset

We used the UMD common activities dataset [26], a dataset of common activities to perform preliminary experiments to validate our model. The dataset consists of 10 activities and 10 different instances of each activity. We partition the dataset into 10 disjoint sets each containing 1 instance of every activity. In order to test the recognition for each set, we first learn the model parameters from the remaining nine sets and then perform recognition for the test sequences. We repeat the process for each of the 10 sets. Thus we ensure that there is no overlap between the training set and the test sequences. Figure 3 shows the 10 X 100 similarity matrix for using the function space algorithm with the uniform distribution on the space of temporal warps. Each column corresponds to a different test sequence while each row corresponds to a different activity. The strongly block diagonal nature of the similarity matrix indicates that the recognition algorithm performs well. In fact, on this database we obtained 100% recognition using both our algorithms.

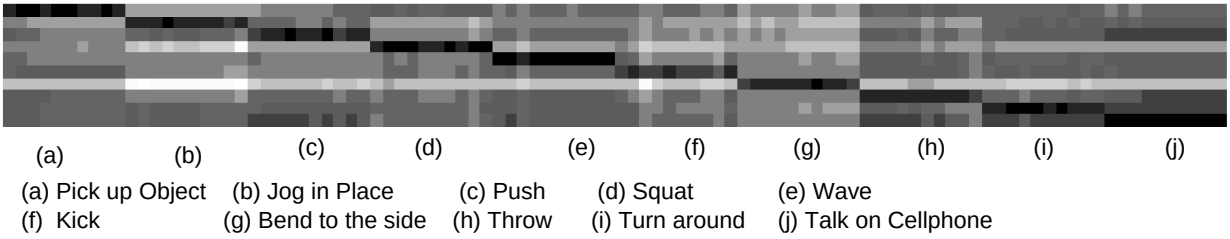


Fig. 3. 10 X 100 Similarity matrix of 100 sequences and 10 different activities using the function space algorithm.

B. INRIA iXmas dataset

The INRIA multiple-camera multiple video database of the PERCEPTION group consists of 11 daily-live motions performed each 3 times by 10 actors. The actors freely change position and orientation. Every execution of the activity is done at a different rate. For this dataset, we extract $16 \times 16 \times 16$ circular FFT features as described in [33]. Since the actors were free to perform the actions the rate at which these actions were performed varied significantly as was shown in Figure 1. So most approaches that cannot handle this vast temporal rate variations, instead model the entire segment as a single motion history volume [33]. Instead, we build a time series of the circular FFT features described in [33]. This allows us to learn the nature of the temporal rate changes between various executions of an action. Using these features, we performed a recognition experiment on the provided data similar to those done in [33]. For the recognition experiment, we used only one segment for each activity which best represented that activity as in [40]. The recognition results are summarized in table I. We used $16 \times 16 \times 16$ circular FFT features in all our experiments here while the results reported in [33] used $32 \times 32 \times 32$ features. The confusion matrix showing confusion between the activities using both the wrapped-Gaussian and the dynamic programming based uniform distribution model are shown in Table II. Note that uniform distribution based model described in Section V is significantly more computationally efficient compared to the Monte-Carlo based inference using the wrapped-Gaussian distribution on the tangent space of warp space.

C. USF Gait Database

Note on gait-based person identification Since the model for learning the function space time-warpings is not explicitly dependent on the choice of features, one could potentially use the same model to learn individual specific function spaces in order to perform activity-based person identification. The only difference would be that we would choose a feature that is

	Activity	**PCA**[33]	**Mahalanobis** [33]	**LDA**[33]	**System Distance** [41]	P_{Unif} (This paper)	P_{Gauss} (This paper)
1	Check Watch	53.33	73.33	76.67	93.33	100	93.33
2	Cross Arms	23.33	86.67	100	100	100	100
3	Scratch Head	46.67	86.67	80	76.67	100	100
4	Sit Down	66.67	93.33	96.67	93.33	96.67	100
5	Get Up	83.33	93.33	93.33	86.67	96.67	100
6	Turn Around	80	96.67	96.67	100	100	100
7	Walk	90	100	100	100	100	100
8	Wave Hand	50	70	73.33	93.33	96.67	96.67
9	Punch	70	86.67	83.33	93.33	83.33	90
10	Kick	50	86.67	90	100	80	100
11	Pick Up	60	90	86.67	96.67	90	100
	Average	61.21	87.57	88.78	93.93	94.85	98.18

TABLE I

COMPARISON OF VIEW INVARIANT RECOGNITION OF ACTIVITIES IN THE INRIA DATASET USING OUR APPROACHES (P_{Unif} AND P_{Gauss}) WITH THE APPROACHES PROPOSED IN [33] AND [41].

Motifs	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Sit Down	30(28)	0(0)	0(1)	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Get Up	0(0)	30(30)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Turn Around	0(0)	0(0)	30(30)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Check Watch	1(0)	0(0)	0(0)	29(30)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Cross Arms	1(0)	0(0)	0(0)	0(0)	29(30)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Scratch Head	0(0)	0(0)	0(0)	0(0)	0(0)	30(30)	0(0)	0(0)	0(0)	0(0)	0(0)
Walk	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	30(30)	0(0)	0(0)	0(0)	0(0)
Wave Hand	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	29(29)	0(0)	0(0)	0(0)
Punch	3(1)	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)	1(1)	25(27)	0(0)	0(0)
Kick	5(0)	0(0)	0(0)	0(0)	1(0)	0(0)	0(0)	0(0)	0(0)	24(30)	0(0)
Pick Up	1(0)	0(0)	0(0)	0(0)	2(0)	0(0)	0(0)	0(0)	0(0)	0(0)	27(30)

TABLE II

CONFUSION MATRIX USING P_{Gauss}(OUTSIDE PARENTHESIS AND P_{Unif} (INSIDE PARANTHESIS) ON THE INRIA DATASET.

person-specific (e.g., silhouette). The nominal activity trajectory would be individual specific in this case. Various external conditions (like surface, shoe) induce systematic time-warping variations within the gait signatures of each individual. The function space of temporal warpings for each individual amounts to learning the class of person specific warping functions. By learning the function space of these variations we are able to account for the effects of such external

conditions. This will allow the same basic approach to be applied for both action recognition and activity based person identification by the use of appropriate features.

In order to compare the performance of our algorithm with the current state of the art algorithms, we also performed a gait-based person identification experiment on the publicly available USF gait database [30]. The USF database consists of 71 people in the Gallery. Various covariates like camera position, shoe type, surface and time were varied in a controlled manner to design a set of challenge experiments[30]. We performed a round-robin recognition experiment in which one of the challenge sets was used as test while the other seven were used as training examples. The process was repeated for each of the seven challenge sets on which results have been reported. Table III [1] shows the identification rates of our algorithm with a uniform distribution on the space of warps (P_{Unif}), our algorithm with a wrapped Gaussian distribution on the tangent space of warps with shape as a feature and with binary image feature (P_{Gauss} and $P_{GaussIm}$). For comparison the table also shows the baseline algorithm [30], simple DTW on shape features [32] and the image-based HMM [31] algorithm on the USF dataset for the 7 probes A-G. Since most of these other algorithms could not account for the systematic variations in time-warping for each class the recognition experiment they performed was not round robin but rather used only one sample per class for learning. Therefore, to ensure a fair comparison, we also implemented a round-robin experiment using the linear warping (P_LW).

The average performance of our algorithms P_{Unif} and P_{Gauss} are better than all the other algorithms that use the same feature, (DTW/HMM (Shape)[32] and Linear warping P_{LW}) and is also better than the baseline[30] and HMM[31] algorithms that use the image as a feature. The image based pHMM algorithm [18] outperforms our algorithm for many probes. One reason for this is that the image as a feature performs better than shape as a feature for the USF dataset. But, it is a computationally very intensive feature (of the order of number of pixels) and consequently leads to algorithms that are very slow. Therefore, we prefer to use the shape as a feature. Inspite of this obvious handicap, the performance of our algorithm is comparable to the image based pHMM algorithm for many probes. The improvement in performance while using binary image as a feature is shown in the last column ($P_{GaussIm}$). The experimental results presented here clearly show that using multiple training samples per class and learning the distribution of their time

[1]Note that the experimental results reported in this table contain varying amounts of training data. While columns 2-6 (Baseline - pHMM) used only the gallery sequences for training, the results reported in columns 7-10 (P_{LW} - $P_{GaussIm}$) used all the probes except the test probe during training.

TABLE III

COMPARISON OF IDENTIFICATION RATES ON THE USF DATASET. NOTE THAT THE EXPERIMENTAL RESULTS REPORTED IN THIS TABLE CONTAIN VARYING AMOUNTS OF TRAINING DATA. WHILE COLUMNS 2-6 (BASELINE - PHMM) USED ONLY THE GALLERY SEQUENCES FOR TRAINING, THE RESULTS REPORTED IN COLUMNS 7-10 (P_{LW} - $P_{GaussIm}$) USED ALL THE PROBES EXCEPT THE TEST PROBE DURING TRAINING.

Probe	Baseline	DTW Shape	HMM Shape	HMM Image	pHMM [18]				P_{LW}	P_{Unif}	P_{Gauss}	$P_{GaussIm}$
Avg.	42	42	41	50	65				51.5	59	59	64
A	79	81	80	96	85				68	70	78	82
B	66	74	72	86	89				51	68	68	78
C	56	52	56	74	72				51	81	82	76
D	29	29	22	32	57				53	40	50	48
E	24	20	20	28	66				46	64	51	54
F	30	19	20	17	46				50	37	42	56
G	10	19	19	21	41				42	53	40	55

warps makes significant improvement to gait recognition results. While most algorithms based on learning from a single sample led to overfitting and therefore performed much better when the gallery was similar to the probe (Probe A-C), they also performed very poorly when the gallery and the probes were significantly different. But, since our algorithm has good generalization ability (becasue we learn the distribution of time warps) the performance of our algorithm did not suffer from overfitting and therefore did not drop as much when moving from probes A-C to Probles D-G.

The importance of using multiple training samples for the problem of gait-based human identification was also recently pointed out in [42]. In order to tackle the lack of training samples, they combine real templates with synthetic templates generated by simulating silhouette distortion. They then develop a statistical spatio-temporal gait representation called Gait Energy Image, that they use in order to perform classification. They show that the generalizing ability afforded by learning from multiple training samples helps gait recognition performance significantly. In our experiments, we used only the available real sequences for training. It might be an interesting alternative to use synthetic training sequences as presented in [42] in cases where the number of available training samples is limited. Recently, an extension of Principal Component analysis for multi-dimensional data called MPCA (Multilinear Principal Component Analysis) was proposed [43]. Such tensor based algorithms for dealing with multi-dimensional data (such as silhouette

sequences) are an important ingredient in performing formal statistical estimation for tensor data. Nevertheless, in its application to gait-based person identification, such algorithms are still limited in their ability to tackle non-linear time warpings since these must first be normalized in a preprocessing step before the gait sequences are converted into tensor data. In this regard, it might be an interesting avenue of further study to combine the non-linear time-warp normalization procedure presented in this paper with statistical tensor analysis approach presented in [43].

VII. Other applications

A. Clustering Activity Sequences

Algorithm for Clustering There are several scenarios where one requires a clustering algorithm to be rate-invariant. Under such scenarios it becomes reasonable to use the rate-invariant model for activities described above as the basis for clustering. When rate-invariance is not a desirable property traditional clustering algorithms such as K-nearest neighbour might be reasonable choices for clustering. We performed clustering experiments on the UMD common activities dataset and the USF gait database using the fast and computationally efficient uniform distribution version of the algorithm denoted by P_{Unif}. The clustering algorithm, based on expectation maximization (EM) is very similar to the Lloyd-Max algorithm [44] and can be used to organize a database of sequences for efficient retrieval. Let us assume that we know the number of clusters, N and the cluster centers $c_1, c_2, ...c_N$. Then, each of the sequences in the database can be associated with one of N clusters. This can be done using a maximum-likelihood approach as described earlier in (25). This forms the Maximization step of the EM algorithm. The Expectation step of the algorithm involves recomputing the new cluster centers from cluster memberships evaluated during the Maximization step. We iterate these 2 steps until convergence. In all our experiments, we initialized the cluster centers randomly.

Clustering on Common Activities Dataset We performed a clustering experiment on the 100 activity sequences collected as a part of the Common Activities dataset. We chose the number of clusters N to be 10 since there were 10 different activities. If clustering were perfect, then the 100 activity sequences would be clustered into 10 different clusters, each cluster containing 10 sequences that correspond to that particular activity. But in reality, clustering would be imperfect and some of the 100 sequences would be misaligned in the wrong cluster. We repeated the clustering experiment several (about 50) times, with a random initialization of cluster centers during each trial. On an average, the algorithm converged in about 10 iterations and about 92%

of the sequences were clustered correctly. Even during some adverse initializations the clustering performance was greater that 80%.

B. Organizing a Large Database of Activities

With the decreasing cost of storage, the size of activity databases is increasing rapidly. For example, the complete USF gait database [30] consists of about 122 classes and a total of more than 1000 sequences. As the size of the database increases, the number of 'distance' computations that must be performed on every query also increases linearly with the size of the database. This poses a significant bottleneck for practical activity recognition systems. We show that organizing the database of sequences using the clustering algorithm described in Section VII-A decreases this computational burden significantly. The price paid is a small decrease in recognition performance. We organize the database of activities in the form of a dendrogram as shown in Figure 4. At each level of the dendrogram the number of branches (B) was set to 3. The number of levels to which the dendrogram is 'grown' determines the trade-off between computation and accuracy. As the number of levels is increased, the number of 'distance' computations that must be performed before finding the class membership of a given test sequence decreases. Therefore, the computational burden of the algorithm also decreases. But this might introduce a decrease in classification performance. When the dendrogram is fully grown (i.e., when each leaf of the dendrogram represents one activity), there will be $log_B N$, levels and therefore $Blog_B N$ 'distance computations'. Let us consider the USF database which consists of 122 subjects and a total of 1870 sequences. A nearest neighbour classifier on this database must perform 1870 distance computations in order to classify a new test sequence. But if we assume that we organize the database in the form of a 'fully grown dendrogram', with each leaf node representing each of the 122 individuals, then one would just have to perform about $Blog_B N = 3 * log_3 122 \approx 14$ 'distance computations'. This is a very significant computational saving.

We performed an experiment to evaluate the efficiency of organizing the database on a subset of the USF database as in Section VI-C. In our experiments, we grow the dendrogram upto 2 levels. We measure efficiency of organization (η) as a ratio of the recognition rate before and after organization.

$$\eta = 100 * \frac{\text{Identification rate after organization}}{\text{Identification rate before organization}} \tag{26}$$

The efficiency η is strongly related to clustering performance and it is reasonable to expect the efficiency η to increase with better clustering. Table IV shows the efficiency of organization for the various probes in the USF dataset. On this data, the dendrogram organization of the database

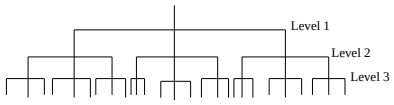


Fig. 4. Dendrogram for organizing an activity database

reduced the computational time by a factor of about 30. This means that the processing time for large databases will be reduced from the order of days to a matter of hours. For such significant reduction in processing time, the Table IV shows that the decrease in recognition performance is not drastic.

TABLE IV
EFFICIENCY OF ORGANIZATION ON THE USF DATASET

Probe	A	B	C	D	E	F	G	Avg
η	76	81	84	100	82	100	95	89

VIII. SUMMARY AND CONCLUSIONS

In this paper, we address an important but often neglected problem in modeling an activity, that of temporal warping of the activity trajectories. Our model for an activity describes each activity using a nominal activity trajectory and a probability distribution on the space of permissible temporal warpings. We discuss the case of a parameteric wrapped-Gaussian distribution on the tangent space of time-warps and derive Monte Carlo sampling-based Bayesian algorithm for classification. We then discuss the spacial case of a convex uniform distribution on the space of time-warps and show that this special case allows us to derive computationally efficient algorithms for a slight decrease in modeling efficieny and classification performance. Finally, we present several experimental results on publicly available action recognition and gait-based person identification datasets.

ACKNOWLEDGMENT

The first author would like to thank Aswin Sankaranarayanan and Pavan Turaga for several discussions during the early development of this work. Many thanks to Daniel Weinland for all the help provided to us with reference to the INRIA iXMAS dataset.

REFERENCES

[1] E. Muybridge, *The Human Figure in Motion*, Dover Publications, 1901.

[2] G. Johansson, "Visual perception of biological motion and a model for its analysis," *PandP*, vol. 14, no. 2, pp. 201–211, 1973.

[3] J.K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[4] C. Cédras and M. Shah, "Motion-based recognition a survey," *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, 1995.

[5] D.M. Gavrila, "The visual analysis of human movement: A survey," *CVIU*, vol. 73, no. 1, pp. 82–98, January 1999.

[6] A.D. Wilson and A.F. Bobick, *Learning visual behavior for gesture analysis*, Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology, 1995.

[7] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," *Computer Vision and Pattern Recognition*, pp. 994–999, 1997.

[8] C. Vogler and D. Metaxas, "ASL recognition based on a coupling between HMMs and 3D motion analysis," *International Conference on Computer Vision*, pp. 363–369, 1998.

[9] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden Markov models," *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1455–1462, 2003.

[10] S.S. Intille and A.F. Bobick, "A framework for recognizing multi-agent action from visual evidence," *AAAI/IAAI*, vol. 99, pp. 518–525, 1999.

[11] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 742–749, 2003.

[12] S. Park and JK Aggarwal, "Recognition of two-person interactions using a hierarchical Bayesian network," *International Multimedia Conference*, pp. 65–76, 2003.

[13] MT Chan, A. Hoogs, R. Bhotika, A. Perera, J. Schmiederer, and G. Doretto, "Joint Recognition of Complex Events and Track Matching," *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006.

[14] B. Laxton, J. Lim, and D. Kriegman, "Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video," *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, 2007.

[15] Y. Sheikh and M. Shah, "Exploring the space of an action for human action recognition," *ICCV*, Oct 2005.

[16] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber, "Automatic symbolic traffic scene analysis using belief networks," *Proceedings 12th National Conference in AI*, pp. 966–972, 1994.

[17] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *International Journal of Computer Vision*, 2002.

[18] Z. Liu and S. Sarkar, "Improved gait recognition by gait dynamics normalization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 863–876, 2006.

[19] A. Veeraraghavan, AK Roy-Chowdhury, and R. Chellappa, "Matching Shape Sequences in Video with Applications in Human Movement Analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 12, pp. 1896–1909, 2005.

[20] A. Kale, AKR Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*, pp. 143–150, 2003.

[21] V. Parameswaran and R. Chellappa, "View invariants for human action recognition," *CVPR*, 2003.

[22] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, 2001.

[23] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, 2005.

[24] A. Gritai, Y. Sheikh, and M. Shah, "On the use of anthropometry in the invariant analysis of human actions," *ICPR*, 2004.

[25] A. Bobick and Tanawongsuwan, "Performance analysis of time-distance gait parameters under different speeds," *4th Intl. Conf. on AVBPA*, June 2003.

[26] A. Veeraraghavan, R. Chellappa, and A.K. Roy-Chowdhury, "The Function Space of an Activity," *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 1*, pp. 959–968, 2006.

[27] L. Rabiner and B. Juang, *Fundamentals of speech recognition.*, Prentice Hall, 1993.

[28] P. Maurel and G. Sapiro, "Dynamic shapes average," *www.ima.umn.edu/preprints/may2003/1924.pdf*.

[29] C.A. Ratanamahatana and E. Keogh, "Making time-series classification more accurate using learned constraints.," *Proceedings of SIAM International Conference on Data Mining*, pp. 11–22, 2004.

[30] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer, "The humanid gait challenge problem: data sets, performance, and analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 162–177, Feb 2005.

[31] A. Kale, A. Sundaresan, A.N. Rajagopalan, N. Cuntoor, A. Roy Cowdhury, V. Krueger, and R. Chellappa, "Identification of humans using gait," *IEEE Trans. on Image Processing*, Sept. 2004.

[32] A. Veeraraghavan, A. RoyChowdhury, and R. Chellappa, "Role of shape and kinematics in human movement analysis," *CVPR*, 2004.

[33] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[34] I.L. Dryden and K.V. Mardia, *Statistical shape analysis*, John Wiley and sons, 1998.

[35] A. Bhattacharya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[36] A. Srivastava, I. Jermyn, and S. H. Joshi, "Riemannian analysis of probability density functions with applications in vision," in *CVPR*, 2007.

[37] Karcher, H., "Riemannian center of mass and mollifier smoothing," *Communications on Pure and Applied Mathematics*, vol. 30, pp. 509–541, 1977.

[38] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu, "Statistical shape analysis: Clustering, learning and testing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 590–602, 2005.

[39] A. Doucet and N. De Freitas, *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[40] D. Weinland, R. Ronfard, and E. Boyer, "Automatic Discovery of Action Taxonomies from Multiple Views," *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 2*, pp. 1639–1645, 2006.

[41] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Unsupervised View and Rate Invariant clustering of Video Sequences," *Computer Vision and Image Understanding*, Accepted for Publication.

[42] J. Han and B. Bhanu, "Individual Recognition Using Gait Energy Image," *IEEE Transactions on Pattern Analysis And Machine Intelligence*, pp. 316–322, 2006.

[43] H. Lu, KN Plataniotis, and A.N. Venetsanopoulos, "MPCA: Multilinear Principal Component Analysis of Tensor Objects," *IEEE Transactions On Neural Networks*, vol. 19, no. 1, pp. 18, 2008.

[44] A.K. Jain, *Fundamentals of digital image processing*, Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1989.