On the Blind Classification of Time Series

Alessandro Bissacco
Google, Inc.
605 Arizona Avenue
Santa Monica, CA 90401
bissacco@gmail.com

Stefano Soatto
Computer Science Department
University of California, Los Angeles
Los Angeles, CA 90095
soatto@cs.ucla.edu

Abstract

We propose a cord distance in the space of dynamical models that takes into account their dynamics, including transients, output maps and input distributions. In data analysis applications, as opposed to control, the input is often not known and is inferred as part of the (blind) identification. So it is an integral part of the model that should be considered when comparing different time series. Previous work on kernel distances between dynamical models assumed either identical or independent inputs. We extend it to arbitrary distributions, highlighting connections with system identification, independent component analysis, and optimal transport. The increased modeling power is demonstrated empirically on gait classification from simple visual features.

1. Introduction

The ability to classify and recognize events as they unfold is an important skill for biological as well as engineering systems to have. The intentions of a predator in the wild, or a suspicious individual at an airport, may not be obvious from its stance or appearance, but may become patent by observing its behavior over time. The pioneering experiments of Johansson [8] with moving dots illustrated eloquently just how much information is encoded in the observation of time series of data, as opposed to static snapshots. Animator artists know how to attach character to inanimate objects by skillfully designing their motion.

Classification and recognition of events is a very complex issue that depends on what kind of sensor data are available (e.g. optical, acoustic), and what representation is used to characterize the events of interest. Certainly a given event (e.g. a traffic accident) can manifest itself in many possible ways, and a suitable representation should exhibit some sort of invariance or insensitivity to nuisance factors (e.g. illumination and viewpoint for the case of optical images) since it is unlikely that one could "train them away" with extensive datasets. Also, many cues contribute to our perception of events. For instance, it is easy to tell that a person is running (as opposed to, say, walking) by a static image snapshot, and whether the classification task is trivial or impossible depends on the null set as well as the alternate hypotheses: It is easy to tell a walking person from a banana even from a static image, but it is not so easy to tell whether she is limping.

But whatever the sensors, whatever the representation, and whatever the null set, in the end one will need the ability to *compare time series of data.*

1.1. From events to dynamical models

If we think of a discrete-time series $\{y(t) \in \mathbb{R}^N\}_{t=1,\dots,T}$ as a function $y : \mathbb{N}_+ \to \mathbb{R}^N$, then comparison between any two sets of data can be performed with any functional norm. However, it will be difficult to do so while discounting simple nuisances such as reparameterizations of the spatial and temporal scale, or the initial time of the experiment: For instance, we may want to recognize a person from her gait regardless of speed, or detect a ball bouncing regardless of height. For this reason, we find it more helpful to think of the a time series as *the output of a dynamical model* driven by some stochastic process. So what we will propose, in the end, will be just another functional norm, but one that is tailored to processes that have dynamic constraints.

Under mild assumptions [9] $y(t)$ can be expressed as an instantaneous function of some "state" vector $x(t) \in \mathbb{R}^n$ that evolves in time according to an ordinary differential equation (ODE) driven by some deterministic or stochastic "input" $v(t)$ with measurement noise $w(t)$, where these two processes are jointly described by a density $q(\cdot)$, which can be degenerate for the limiting case of deterministic inputs. They can be thought of as errors that compound the effects of unmodeled dynamics, linearization residuals, calibration errors and sensor noise. For this reason they are often collectively called input (state) and output (measurement) "noises." For reasons that will become clear shortly,

1-4244-1180-7/07/$25.00 ©2007 IEEE

we assume that the noise process is temporally independent, or *strongly white.* In general, $q(\cdot)$ may or may not be Normal. For the case of human gaits, one can think of limit cycles generating nominal input trajectories,[1] stable dynamics governing muscle masses and activations, initial conditions characterizing the spatial distribution of joints, and the input depending on the actual gait, the terrain, and the neuromuscular characteristics of the individual.

So, comparing time series entails *endowing the space of dynamical models with a metric structure,* so we can measure the distance between models. Such a distance should include all elements of the model: The input, the state and its dynamics, the output map, the initial condition, but allow the possibility of discounting some depending on the task at hand.

The simplest conceivable class of dynamical models is linear ones, where the time series $\{y(t)\}$ is generated via the model

$$\begin{cases} x(t+1) = Ax(t) + v(t) \quad x(t_0) = x_0 \\ y(t) = Cx(t) + w(t) \quad \{v(t), w(t)\} \overset{IID}{\sim} q(\cdot) \end{cases} \tag{1}$$

that is determined by the matrices $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times n}$, and the density $q(\cdot)$. If the latter is Gaussian with zero-mean, this is determined by its covariance $Q \in \mathbb{R}^{k\times k}$.

1.2. Comparing dynamical models

Consider the simplest model (1), driven by a Gaussian input. It can be thought of as a point in a space that is embedded in $\mathbb{R}^{n^2+mn+k^2}$. Unfortunately, even for such linear-Gaussian models, this space is non-linear, due to the constraints on the parameters and the fact that there are equivalence classes of models that yield the same output statistics [12]. Thus, in order to define a proper distance that takes into account the geometry of the space, one would have to define a Riemannian metric in the (homogeneous) space of models and integrate it to find geodesic distances between any two points on the space.

A simpler approach is to define a *cord distance*, one that does not come from a Riemannian metric, between any two points. Many such distances have been proposed recently, from subspace angles [5] to spectral norms [15], to kernel-induced distances [21].

1.3. Linear-Gaussian is not enough

Linear-Gaussian models capture the second-order statistics of a stationary time series. In fact, there is an entire equivalence class of models that have the same second-order statistics, so it is common to choose a realization that is stable and has minimum phase as a representative of the class [13, 19, 16]. That is often sufficient to classify coarse classes of events [2].

However, the assumptions of stability and phase minimality are often violated in important classes of data: For instance, humans are a collection of inverted penduli, the prototypical example of non-minimum phase mechanical systems [9], and their gaits are often quasi-periodic, or marginally stable. So, we need to broaden our attention to non-minimum phase, marginally stable models.

Unfortunately, the moment we allow non-minimum phase behavior, unstable zeros of the system allow higher-order statistics in the input through [16]. So it is possible that models that have identical parameters but input distributions that differ by higher-order statistics produce different output time series [4, 20]. This forces us to *model the higher-order statistics of the input,* and forego the Gaussian assumption. The linearity assumption, on the other hand, is less limiting since many non-linear dynamics can be embedded into higher-dimensional linear models [11], barring hysteresis, turbulence and other intrinsically non-linear phenomena.

So, the class of models we are honing into is linear ones, possibly of unknown order, with non-Gaussian input. Since one can interpret a white non-Gaussian independent and identically distributed (IID) process as a Gaussian one filtered through a static non-linearity, we are left with considering so-called *Hammerstein models*, that are linear models with static input non-linearities [6]. In [3] we have shown how to perform identification (learning) of such models, so we will not address the learning part here. Instead, we will concentrate on the problem of endowing Hammerstein systems with a cord distance.

1.4. What we propose in this paper

Our starting point is the work of Smola et al. [21], that introduce an inner product in the embedding space of an output time series and use it to define a cord distance between dynamical models. Although their derivation is elegant and general, in order to compare two models M_1, M_2, the method proposed in [21] requires knowledge of the *joint density* of the noises, i.e. $p(v_1, w_1, v_2, w_2)$, which is seldom available. *We propose a novel distance between dynamical models that can be computed without such knowledge.*

The main idea of our method is to *identify a model that generates the same output statistics (of all orders) of the original system, but that has a canonical input* that is strongly white and with independent components. Then all the information content of the input is transferred to the model, that becomes non-linear (Hammerstein) even if the original one was linear. One can then proceed to define a kernel in a manner similar to [21], but extended to take into account the non-linearity. This can be done by solving an *optimal transport* problem which, given a finite amount of

[1] This input can itself be considered as the output of an "exo-system," for instance the central nervous system, that is not explicitly modeled.

data, can be done in closed-form.

Thus completing the work of Smola et al. will require us to explore links with independent component analysis (ICA), and optimal transport (Wasserstein).

1.5. Learning preliminaries

While in this manuscript we do not address the issue of learning the model (1), we summarize the procedure used in [3] to convert it, without losing generality, into a form that will allow us to easily define and compute a kernel. The first step is to rewrite the model in innovation form

$$\begin{cases} x(t+1) = Ax(t) + Kn(t) \\ y(t) = Cx(t) + n(t). \end{cases} \tag{2}$$

Under our assumptions the noise $n(t)$ is temporally (strongly) white, and its components are *weakly* independent (uncorrelated). Then we can *normalize* this model to make the components of the noise *strongly independent*. This is equivalent of performing *independent component analysis* (ICA) $n(t) = D\epsilon(t)$, yielding a model of the form

$$\begin{cases} x(t+1) = Ax(t) + B\epsilon(t) \\ y(t) = Cx(t) + D\epsilon(t) \end{cases} \tag{3}$$

with $B = KD$ and the components of ϵ are independent zero-mean unit-variance IID processes

$$\begin{aligned} \epsilon(t) &= \begin{bmatrix} \epsilon_1(t) & \epsilon_2(t) & \cdots & \epsilon_m(t) \end{bmatrix}^\top \\ \epsilon_i(t) &\overset{IID}{\sim} q_i(\epsilon_i) \quad , \quad E[\epsilon(t)\epsilon(t)^\top] = I \end{aligned} \tag{4}$$

and ϵ can be written in terms of a canonical (e.g. uniform, or Gaussian) noise u:

$$\begin{cases} x(t+1) = Ax(t) + Bf(u(t)) \quad x(t_0) = x_0 \\ y(t) = Cx(t) + Df(u(t)). \end{cases} \tag{5}$$

It is on this representation of the model $M = \{A, B, C, D, x_0, f\}$ that we define a kernel, and therefore a distance, that takes into account the dynamics, the measurement map, the initial conditions, *and* the input statistics. The hypothetical experiment to compare two models consists of randomly generating a scalar IID sequence distributed uniformly in $[0\ \ 1]$, feeding it to the two models, and then compare their outputs; see [3], Sections 2 and 3, for more details.

2. Kernels for Linear Systems

In this section we will define kernels for dynamical systems of the form $(3, 4)$ with input $\epsilon(t) \in \mathbb{R}^m$, state $x(t) \in \mathbb{R}^n$ and output $y(t) \in \mathbb{R}^m$. Here we only assume that the input is a unit variance IID stationary process with independent components. In the next section we will complete the model $(3, 4)$ to include the higher-order statistics of the process $y(t)$ by explicitly representing the distribution of the input components $\epsilon_i(t)$.

Given two linear models $M = \{A, B, C, D, x_0\}$, $M' = \{A', B', C', D', x_0'\}$ and the unit-variance inputs $\epsilon(t), \epsilon'(t)$, we obtain the following outputs $y(t), y'(t)$:

$$\begin{aligned} y(t) &= CA^t x_0 + D\epsilon(t) + \sum_{i=0}^{t-1} CA^{t-1-i} B\epsilon(i) \\ y'(t) &= C'(A')^t x_0' + D'\epsilon'(t) + \sum_{i=0}^{t-1} C'(A')^{t-1-i} B'\epsilon'(i) \end{aligned} \tag{6}$$

If the inputs were Gaussian or had the same higher-order statistics, we could define kernels between models (7) by assuming the same input:

$$\epsilon'(t) = \epsilon(t). \tag{7}$$

This allows us to compute the correlation matrix Σ between $y(t)$ and $y(t)'$ by marginalizing over the common noise $\epsilon(t)$:

$$\Sigma\left[M, M'\right] \doteq E_\epsilon \left[\sum_{t=1}^{\infty} e^{-\lambda t} W y'(t) y(t)^\top \right] \tag{8}$$

where, following [21], we use an exponential discounting factor $e^{-\lambda t}$, $\lambda \geq 0$ and a user-defined symmetric weight matrix W. From (7) we have:

$$\begin{aligned} \Sigma\left[M, M'\right] &= \Sigma\left[\{A, C, x_0\}, \{A', C', x_0'\}\right] + \\ &+ \Sigma\left[\{A, B, C, D\}, \{A', B', C', D'\}\right] \end{aligned} \tag{9}$$

where:

$$\Sigma\left[\{A, C, x_0\}, \{A', C', x_0'\}\right] = \sum_{t=1}^{\infty} e^{-\lambda t} W C' (A')^t x_0' x_0^\top (A^\top)^t C^\top \tag{10}$$

$$\begin{aligned} \Sigma\left[\{A,B,C,D\},\{A',B',C',D'\}\right] = E_\epsilon \Bigg[\sum_{t=1}^{\infty} e^{-\lambda t} W \Big(D'\epsilon'(t)\epsilon(t)^\top D^\top + \\ + \sum_{i=0}^{t-1} C'(A')^{t-1-i} B'\epsilon(i)'\epsilon(i)^\top B^\top (A^\top)^{t-1-i} C^\top \Big) \Bigg] \end{aligned} \tag{11}$$

The correlation on the initial state (10) can be computed as in [21]:

$$\begin{aligned} \Sigma[\{A, C, x_0\}, \{A', C', x_0'\}] &= WC'VC^\top \\ V &= e^{-\lambda} A' x_0' x_0^\top A^\top + e^{-\lambda} A' V A^\top \end{aligned} \tag{12}$$

The correlation on the noise (11) is:

$$\begin{aligned} \Sigma\left[\{A, B, C, D\}, \{A', B', C', D'\}\right] &= (e^\lambda - 1)^{-1} W \left(D'UD^\top + C'\tilde{V}C^\top \right) \\ \tilde{V} &= B'UB^\top + e^{-\lambda} A' \tilde{V} A^\top \end{aligned} \tag{13}$$

where $U \doteq E_\epsilon[\epsilon'(t)\epsilon(t)^\top]$. Now, if we assume the noises have the same input (7) and have unit variance (4), then

$U = I$. However, we are interested in extending this to arbitrary distributions, and in the next subsection we will use the input correlation matrix U to include the effect of the higher-order statistics of the input distributions. Then, from the output correlation matrix (8), we can define the trace kernel k_t as:

$$k_t(M, M') \doteq E_\epsilon \left[\sum_{t=1}^{\infty} e^{-\lambda t} y(t)^\top W y'(t) \right] = \quad (14)$$

$$= \text{tr}\Sigma \left[M, M' \right] \quad (15)$$

and the determinant kernel k_d as:

$$k_d(M, M') \doteq E_\epsilon \text{det} \left[\sum_{t=1}^{\infty} e^{-\lambda t} y'(t) y(t)^\top \right] =$$

$$= \text{det}\Sigma \left[\{A, B, C, D, x_0\}, \{A', B', C', D', x_0'\} \right] \quad (16)$$

where, without loss of generality, we assume det$W = 1$. Using the Binet-Cauchy theorem on compound matrices, in [21] it is shown that functions of the form $(15, 16)$ are inner products in an embedding space and they define positive definite kernels.

The trace kernels (15) provide several computational and theoretical advantages over determinant kernels (16)[2]. Therefore in the extensions that follow we will consider trace kernels alone.

The proposed kernels can be used to define a distance in the space of linear models. Let $M = \{A, B, C, D, x_0\}$, $M' = \{A', B', C', D', x_0'\}$ be two such models, then the kernel distance $d(M, M')$ is defined as:

$$d(M, M')^2 = k(M, M) + k(M', M') - 2k(M, M') \quad (17)$$

This is a crucial ingredient to perform classification in the space of dynamical models.

2.1. Kernels for Arbitrary Input Distributions

In this section we will introduce the last necessary element of our approach, a kernel between arbitrary IID processes. Given a random variable x with density function p and cumulative distribution function $F : \mathbb{R} \mapsto [0, 1]$:

$$x \sim p(x) \quad , \quad F(a) = \int_{-\infty}^{a} p(x)dx = P[x \leq a] \quad (18)$$

we can use the quantile function F^{-1} (i.e. the inverse of the distribution function) to transform a uniform variate $u \in \mathbb{U}[0, 1]$ into a random variable distributed according to F:

$$u \in \mathbb{U}[0, 1] \ \rightarrow \ F^{-1}(u) \sim p(x). \quad (19)$$

Thus, we can define a kernel between pairs of (scalar) random variables x, x' having distributions F, F' as the correlation between the two random variables obtained by applying the same uniform u to the quantile functions F^{-1}, F'^{-1}:

$$k(x, x') = E_{u \sim \mathbb{U}[0,1]}[F^{-1}(u) F'^{-1}(u)] = \int_0^1 F^{-1}(u) F'^{-1}(u) du \quad (20)$$

Consider the linear manifold[3] $\mathcal{H}$ of random variables with zero mean and finite variance defined on the same probability space $(\Omega, \mathcal{F}, P)$. It is well known [17] that $\mathcal{H}$ can be made into an Hilbert space introducing the inner product $\langle x, x' \rangle \doteq E[xx']$.

Then, (20) is an inner product and consequently a positive definite kernel. The distance induced by this kernel:

$$d_W(x, x')^2 = k(x, x) + k(x', x') - 2k(x, x')$$

$$= \int_0^1 |F^{-1}(u) - F'^{-1}(u)|^2 du \quad (21)$$

is known for probability distributions as Wasserstein, Mallows or Ornstein distance [14, 1]. It is more generally defined for two (possibly multidimensional) probability densities P and Q as $d_W(P, Q)^2 = \inf_J \{E_J[(X - Y)^\top (X - Y)] : \ (X, Y) \sim J, X \sim P, Y \sim Q\}$, where the infimum is taken over all the joint densities J which have marginals equal to P and Q. This distance represents the solution to the Monge-Kantorovich mass transfer problem, and can be interpreted as the minimum amount of work that is required to transport a mass of soil with distribution P to an excavation having distribution Q. For discrete distributions, the Wasserstein distance is equivalent to the Earth mover's distance, a metric commonly used for measuring texture and color similarities.

From (21), we can compute the kernel between input distributions $k(x, x')$ from their Wasserstein distance $d_W(x, x')$. Using the change of variable $x = F^{-1}(u)$, it is easy to see that the kernel $k(x, x)$ gives the second moment of x:

$$k(x, x) = \int_0^1 |F^{-1}(u)|^2 du = \int_{-\infty}^{\infty} x^2 p(x) dx = E[x^2]. \quad (22)$$

Substituting (22) in (21) we obtain:

$$k(x, x') = \frac{1}{2} \left(E[x^2] + E[x'^2] - d_W(x, x') \right) \quad (23)$$

[2] First they allow for more efficient computations in the case of high-dimensional data, since they can be computed from a $n \times n$ matrix derived from the inner product (14) instead of the determinant kernel which need to use the high-dimensional correlation matrix (10) (see [21] for details on calculations). When the measurements $y(t)$ are images, trace kernels are indeed the only computationally doable option. Another advantage of trace kernels compared to determinant kernels is that they do not introduce ambiguities on the sign of the correlation. For example if $y(t)$ has an even number of independent components and $y'(t) = -y(t)$, then the determinant kernel will give the same score as when the two processes are the same, while the trace kernel correctly identifies their negative correlation. Finally, the linearity of trace kernels allows us to decompose the final result as the sum of the single contributions, that is initial state evolution (12) and input distribution (13).

[3] I.e. the space of finite linear combinations of random variable in $(\Omega, \mathcal{F}, P)$, closed with respect to convergence in mean square.

In case of zero-mean unit variance $E[x^2] = E[x'^2] = 1$, we have simply $k(x, x') = 1 - \frac{1}{2} d_W(x, x')^2$.

Although this expression is attractive, in the case of discrete distributions it is more efficient to compute the kernel by directly evaluating the integral (20). We can define a kernel between scalar IID processes $x(t), x'(t)$ as:

$$\begin{aligned} k(x(t), x'(t)) &\doteq E_u \left[\sum_{t=1}^{\infty} e^{-\lambda t} F^{-1}(u(t)) F'^{-1}(u(t)) \right] \\ &= (e^{\lambda} - 1)^{-1} E_u \left[F^{-1}(u) F'^{-1}(u) \right]. \quad (24) \end{aligned}$$

Now we extend the kernel (24) to multivariate processes. Given an IID process $\epsilon(t) \in \mathbb{R}^m$ with independent components, it can be modeled as the output of its m quantile functions F_i^{-1} to m independent uniform processes $u_i(t)$, i.e. $\epsilon(t) = f(u(t))$, where $f(u(t)) = \begin{bmatrix} F_1^{-1}(u_1(t)) & \cdots & F_m^{-1}(u_m(t)) \end{bmatrix}$. Then, given two IID processes $\epsilon(t), \epsilon'(t) \in \mathbb{R}^m$ with independent components, they can be represented as outputs of two vector functions f, f' to the same input u:

$$\epsilon(t) = f(u(t)) \quad , \quad \epsilon'(t) = f'(\Pi(\sigma) u(t)) \qquad (25)$$

where $\sigma \in S(m)$ (symmetric group of order m) is a permutation of the input representing correspondences between the elements of the two processes, i.e. each component i of ϵ is correlated with the component σ_i of ϵ' : $E[\epsilon_i \epsilon'_{\sigma_i}] \neq 0$, $E[\epsilon_i \epsilon'_j] = 0$ $j \neq \sigma_i$, and $\Pi(\sigma) = [\pi_{ij}]$ is the permutation matrix corresponding to σ, i.e. $\pi_{i\sigma_i} = 1$, $\pi_{ij} = 0 \; \forall j \neq \sigma_i$.

If the processes $\epsilon(t), \epsilon'(t)$ are inputs to a linear model of the form (3), the permutation σ represents the inherent ambiguity of the model, since we can obtain equivalent systems by rearranging the input elements $\epsilon_i(t)$ and the columns of the mixing matrix D. Additionally, there is a sign ambiguity, that is we can change the sign of any $\epsilon_i(t)$ and of the corresponding i-th column of D.

Using (25), we can compute the correlation matrix U between vector processes $\epsilon(t), \epsilon'(t)$ with correspondences σ as:

$$\begin{aligned} U(\sigma) &\doteq E_u \left[\sum_{t=1}^{\infty} e^{-\lambda t} f'(\Pi(\sigma) u(t)) f(u(t))^{\top} \right] = \qquad (26) \\ &= \Pi(\sigma) \begin{bmatrix} k(\epsilon_1(t), \epsilon'_{\sigma_1}(t)) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k(\epsilon_m(t), \epsilon'_{\sigma_m}(t)) \end{bmatrix} \end{aligned}$$

Given the correspondences σ, we can define the trace kernel between $\epsilon(t)$ and $\epsilon'(t)$ as:

$$k_t(\epsilon(t), \epsilon'(t); \sigma) = \text{tr}\left(\Pi(\sigma)^{\top} |U(\sigma)| \right) \sum_{i=1}^{m} |k\left(\epsilon_i(t), \epsilon'_{\sigma_i}(t)\right)| \qquad (27)$$

where we use the absolute value of the correlation between input components to resolve the sign ambiguity. This is a symmetric positive function of the input distributions, therefore is a positive definite kernel [18]. If the correspondences σ are unknown, we can compute the optimal trace matching $\hat{\sigma}_t$ as the solution to the maximum-weight assignment problem defined by the $m \times m$ Gram matrix $\mathcal{K} = \left[|k(\epsilon_i(t), \epsilon'_j(t))| \right]$:

$$\begin{aligned} \hat{\sigma}_t &\doteq \arg \max_{\sigma \in S(m)} k_t(\epsilon(t), \epsilon'(t); \sigma) = \\ &= \arg \max_{\sigma \in S(m)} \sum_{i=1}^{m} \left| k(\epsilon_i(t), \epsilon'_{\sigma_i}(t)) \right|. \qquad (28) \end{aligned}$$

The optimal matching problem (28) can be solved in $O(m^3)$ using the Hungarian algorithm [10]. We use these results to extend the trace kernels between linear systems (15) to include the effect of the input distributions. To do so, we apply the correlation matrix $U(\sigma)$ given in (27) in the calculation of the noise related matrix $\Sigma[\{A, B, C, D\}, \{A', B', C', D'\}]$ (13). In particular, we apply the correlation $U(\hat{\sigma}_t)$ corresponding to the optimal assignment $\hat{\sigma}_t$ solution to the additive matching problem (28). A similar extension can be applied to determinant kernels (16).

3. Experiments

In this section we present results on the applications of the proposed kernels for non-Gaussian systems (15) to the problem of classifying human gaits. These experiments are based on the CMU Mobo dataset [7]. The goal is to identify the 4 classes of walking motions (normal walk, fast walk, walk with ball and walk on inclined treadmill) performed by the 24 subjects in the dataset. We use only the sequences taken from the same viewpoint (camera $vr03\_7$). Each sequence is 340 frames long and is pre-processed to yield simple features using the silhouette. We further coarsen the binary silhouettes and compute Hu moments and PCA to further reduce the dimensionality [3].

In Fig. 1 we show a sample image from background subtraction and the corresponding representation with the projection features. Given a binary silhouette, the projection features encode the distance of the points on the silhouette from lines passing through its center of mass. The bounding box of the silhouette is divided uniformly in $2n$ region, n to each side of the projection line, and for each region the average distance from the line is computed. In our experiments we used 2 lines (horizontal and vertical) and $n = 8$ features on both side, for a total 32 components (Fig. 1).

On the feature trajectories extracted from a video sequence, we apply the learning algorithm proposed in [3] to estimate the parameters of the linear non-Gaussian model (5). As before, in order to obtain better estimates it is advisable to reduce the dimensionality of the data by PCA projection, here we use $m = 8$ components.

For each learned model pair in the dataset we then proceed to compute the full trace kernels (15). These are made of two terms: The similarity between the deterministic part of the systems encoded in periodic components and initial states (12), and the matching between the stochastic parts, represented by kernels on input correlation (13). In Fig. 2 we plot the confusion matrices showing the distances (17) between learned models defined by initial state trace kernels (left) and the full trace kernels, including input distributions (right). It is evident that the inclusion of the stochastic part modeled by the input statistics improves the gait discrimination performances, visible by the block diagonal structure of the corresponding confusion matrix and the higher number of same-gait nearest neighbor matches.

4. Discussion

We have found that linear systems driven by non-Gaussian inputs are a rich-enough class of models for many events of interest in computer vision. Defining a distance between two such models has not been done before. We extend [21] to arbitrary non-Gaussian inputs. This is made possible by a learning (identification) procedure that transforms, without loss of generality, a linear model into one with strongly white inputs with independent components with static non-linearities. We then extend the kernel to these models by defining a component kernel between two inputs, computed by solving an optimal transport problem. The resulting kernel allows the user to discount, depending on the application, the transients, the inputs, or output maps.

5. Acknowledgement

This research was sponsored by ONR N00014-03-1-0850:P0001, AFOSR FA9550-06-0138/E-16-V91-G2 and NSF ECS-0622245.

References

[1] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, Vol. 9, No. 6:1196–1217, 1981.

[2] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto. Recognition of human gaits. In *Proc. CVPR*, pages II 52–58, December 2001.

[3] A. Bissacco, A. Chiuso, and S. Soatto. Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport. *To appear in PAMI*, 2007.

[4] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Wiley, 2003.

[5] K. D. Coch and B. D. Moor. Subspace angles and distances between arma models. *Proc. of the Intl. Symp. of Math. Theory of Networks and Systems*, 2000.

[6] I. Goethals, K. Pelckmans, J. A. K. Suykens, and B. D. Moor. Subspace identification of hammerstein systems using least squares support vector machines. *IEEE Trans. on Automatic Control*, Vol. 50, No. 10:1509–1519, Oct. 2005.

[7] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical report, Robotics Institute, Carnegie Mellon University, 2001.

[8] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[9] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, 2000.

[10] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83, 1955.

[11] J. Levine. Finite dimensional filters for a class of nonlinear systems and immersion in a linear system. *SIAM J. Control and Optimization*, 25(6):1430–1439, 1987.

[12] A. Lindquist and G. Picci. A geometric approach to modelling and estimation of linear stochastic systems. *Journal of Mathematical Systems, Estimation and Control*, 1:241–333, 1991.

[13] L. Ljung. *System Identification; Theory for the User*. Prentice Hall, 1997.

[14] C. L. Mallows. A note on asymptotic joint normality. *Ann. of Mathematical Statistics*, 43:508–515, 1972.

[15] R. Martin. A metric for arma processes. *IEEE Trans. on Signal Processing*, 48(4):1164–1170, 2000.

[16] P. V. Overschee and B. D. Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer, 1996.

[17] Y. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.

[18] B. Schoelkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, 2002.

[19] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, 1989.

[20] A. Swami, G. Giannakis, and S. Shamsunder. Multichannel arma processes. *IEEE Trans. on Signal Processing*, 42(4):898–913, 1994.

[21] S. Vishwanathan, R. Vidal, and A. J. Smola. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *International Journal of Computer Vision*, 2005.

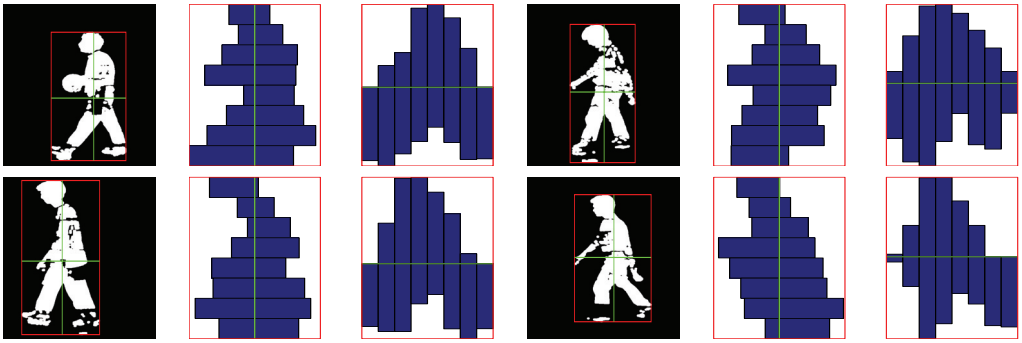

Figure 1. *Sample silhouettes and associated shape features from [7]: walking with ball, normal walk, fast walk and inclined walk. Superimposed to the binary silhouette we plot the bounding box (red) and the horizontal and vertical lines passing through the center of mass used to extract the features. On columns* $(2, 5)$ *and* $(3, 6)$ *we show the features obtained by computing the distance of the points on the two sides of the silhouette to respectively the vertical and horizontal lines, discretized to* $n_f = 8$ *values.*

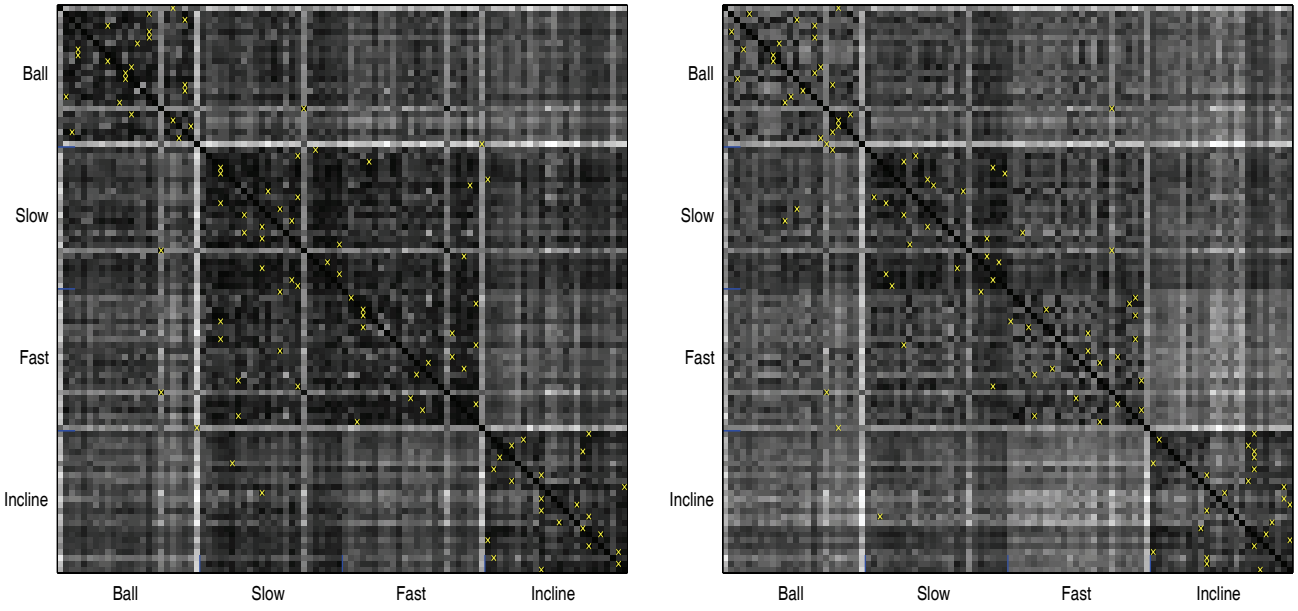

Figure 2. *State and input kernel distances. We show the confusion matrices representing trace kernel distances between non-Gaussian linear models learned from walking sequences in the Mobo dataset. There are* 4 *motion classes and* 24 *individuals performing these motions, for a total of* 96 *sequences. For each sequence we learn a linear model* (5) *and then measure distance between models by the trace kernels. On the left we show results using kernels on initial states only, on the right we display the confusion matrix obtained from the trace kernels that include the effect of the input* (15)*. For each row a cross indicates the nearest neighbor. It is clear how the additional information provided by the input statistics results in improved gait classification performances: we have* 17 (17.7%) *nearest neighbors mismatches (i.e. closest models that do not belong to the same gait class) using the state-only distance, while only* 9 (9.3%) *with the complete trace kernel distance.*