Title
Hybrid Dynamical Models of Human Motion for the Recognition of Human Gaits

Journal
International Journal of Computer Vision, 85(1)

ISSN
1573-1405

Authors
Bissacco, Alessandro
Soatto, Stefano

Publication Date
2009-10-01

DOI
10.1007/s11263-009-0248-7

Peer reviewed

Int J Comput Vis (2009) 85: 101–114
DOI 10.1007/s11263-009-0248-7

Hybrid Dynamical Models of Human Motion for the Recognition of Human Gaits

Alessandro Bissacco · Stefano Soatto

Received: 27 November 2006 / Accepted: 28 April 2009 / Published online: 12 May 2009

Abstract We propose a hybrid dynamical model of human motion and develop a classification algorithm for the purpose of analysis and recognition. We assume that some temporal statistics are extracted from the images, and use them to infer a dynamical model that explicitly represents ground contact events. Such events correspond to "switches" between symmetric sets of hidden parameters in an autoregressive model. We propose novel algorithms to estimate switches and model parameters, and develop a distance between such models that explicitly factors out exogenous inputs that are not unique to an individual or his/her gait. We show that such a distance is more discriminative than the distance between simple linear systems for the task of gait recognition.

Keywords Human motion estimation · Hybrid system identification · Tracking · Gait analysis · Synthesis · Recognition · Dynamical models

1 Introduction

The analysis of human motion has been a subject of interest in the vision community for decades, further reinforced in recent years by applications in security, biomechanics and entertainment. All aspects of the problem, from modeling to detection, tracking, classification, and recognition are the subject of active research (Gavrila 1999; Shah and Jain 1999). From a modeling perspective, humans are physical objects interacting with physical space in ways that are mediated by forces, masses and inertias that can be described, to first approximation, by ordinary differential equations. In other words, humans are dynamical systems. Analytically, each individual can be described by a model that includes intrinsic parameters (masses, inertias), internal *states* (skeletal configurations, internal forces), also a property of the individual, and external forces (*inputs*), including contact forces, that depend on the environment and other nuisance factors. From the point of view of perception, humans and their clothes interact with light and an imaging device to yield *output* images.

While "static" (e.g. pose, skeletal configurations Lee and Grimson 2002), "quasi-static" (e.g. graphs of transitions between poses Sarkar et al. 2005, cumulative video statistics Bobick and Davis 2001), or "kinematic" representations (Bregler 1997) already contain significant information on both the identity of humans and their action,[1] *dynamics* also play a crucial role, that has been recognized early on by Johansson (1973) who showed that even if we strip the images of all of their pictorial content and look at displays of moving dots, from their motion we can often tell whether a person is young or old, happy or sad, man or woman.[2] In this

Research supported by ONR 67F-1080868 and AFOSR FA9550-06-1-0138.

A. Bissacco
Google, INC, 604 Arizona Ave., Santa Monica, CA 90401, USA

S. Soatto (✉)
University of California Los Angeles, 3531D Boelter Hall, 405 Hilgard Avenue, Los Angeles, CA 90095, USA
e-mail: soatto@ucla.edu

[1] It is often easy to tell that someone is running, rather than walking, from a single snapshot.

[2] One could argue that moving dot displays also contain pose and kinematic information; however, dynamics remains an important cue, as one can guess by watching two-hundred pound actors imitating Charlie Chaplin's walk (different masses, inertias and skeletal configuration, same perceived dynamic characteristics). Furthermore, one single snapshot of such moving dot displays rarely yields much information.

paper we concentrate on *dynamics as a perceptual cue for human motion recognition*. This does not mean that kinematics, or pose or even pictorial cues are not important, and eventually all will have to be integrated into a coherent system. We believe, however, that dynamics has been largely unexploited, hence our emphasis in this paper.

If we agree in viewing humans as dynamical systems, then learning their dynamic characteristics is a system identification task (Ljung 1987). System identification is a well established field, and yet in almost 50 years of research the problem of performing decision tasks, such as detection and recognition, in the space of dynamical models is largely unexplored. Several attempts have been made to endow the space of dynamical models with a metric and probabilistic structure, such as the Gap metric (Zames and El-Sakkary 1980), subspace angles (De Coch and De Moor 2000), Martin's distance (Martin 2000). However, even for simple linear systems deciding "how far" two models are is not straightforward, and learning a distribution (e.g. a prior) in model space is even less so (Krishnaprasad and Brockett 1979).[3] In particular, if we want to be able to learn models that have discriminative power, we have to factor out nuisance factors, such as external forces, that do not depend on the particular individual or gait. Therefore, in this work we consider *models that explicitly represent contact dynamics*; such models are *hybrid*, in the sense that they involve both continuous dynamics and discrete "switches." Therefore, the simplest instance of our problem involves performing *inference and classification of hybrid dynamical models.* Since the analysis is complex enough for *repetitive gaits* (e.g. walking, running, jumping), we concentrate on this case. Ideally an individual should be recognized regardless of the gait, and in particular during transient maneuvers, but this is beyond the scope of this paper.

In order to distill the essence of the problem, we concentrate on dynamics, and assume that some representation of a human gait has been inferred, either in the form of joint angles in a skeletal model (e.g. Bregler and Malik 1998), or in the form of joint positions, e.g. from a motion-capture system. In other words, we use data similar to Johansson's displays, that distill dynamic information. Note that, although we assume that the "image-to-model" problem is solved, which is not quite the case even today, and although we do not use any images in this work, the models we study are designed and analyzed for the purpose of vision-based classification: If we were to infer and analyze models for, say, computer graphics, or robotics, or biomechanics, the models would be quite different, and their inference would likely entail additional measurements (e.g. forces) that are not directly available in a vision context. So, we concentrate on *inference and classification of hybrid dynamical models designed for vision-based human motion analysis and recognition.* This is not a trivial problem, and even some of the basic ingredients are missing from the literature, as we explain in the following section.

[3]Note that each of these techniques has been applied to the analysis and classification of human motion (Bissacco et al. 2001 for subspace angles and Martin's distance, Mazzaro et al. 2002 for the Gap metric) with encouraging but limited results.

1.1 Relation to Previous Work

The problem of recognizing human activities compounds several aspects including *modeling* and *inference*. Modeling, in turn, requires addressing the photometric, geometric and dynamic aspects of the image formation process. Photometric modeling addresses the variability in the images due to the interaction of light with matter, specifically clothing for the case of humans. Geometric modeling addresses the variability in the shape of objects in the scene, for instance the pose and deformation of the human figure. Whatever representation of the photometry and geometry of the object of interest, dynamic modeling addresses the temporal variability of such a representation. Naturally there is interplay between these factors, as one can explain the data with infinitely many combinations of different photometric, geometric and dynamic configurations. In this paper, we focus on the dynamic aspect of the problem, and therefore we wish to isolate it as much as possible from the photometric and geometric aspects. For this reason, we use motion capture data, as an abstraction of a representation where the photometric and geometric aspects of the image formation process have been factored out, similarly to what Johansson did for his psychophysical experiments.

Once a model is in place, a number of *inference* techniques can be exploited to estimate its state or identify its parameters. Although general techniques exist to approximate the posterior distribution of the state of any dynamical model (North et al. 2000), they do not exploit the particular structure of our model and are, in this sense, overkill. We propose an inference technique that is tailored to the class of model we have introduced.

In order to justify and validate a particular model, one can consider a variety of end-tasks, for instance the classification of gaits (Shah and Jain 1999) regardless of the individual, or the identification of people regardless of their gait (Sarkar et al. 2005). Most of the approaches in the literature can be classified as either model-based (Bregler 1997; Bissacco et al. 2001; Lee and Elgammal 2004; Kale et al. 2004), whereby motion is represented by parameters in a model within a chosen class, or holistic (Little and Boyd 1998; Veres et al. 2004), where some statistics are extracted from the video sequence and used for classification. In all cases the first step consists in deriving a compact representation of the motion, such as binary silhouettes

(Sarkar et al. 2005; Kale et al. 2004), optical flow (Little and Boyd 1998), joint angles of an articulated body model with image-based tracking (Bregler 1997; Bissacco et al. 2001; North et al. 2000), or other spatio-temporal motion descriptors (BenAbdelkader et al. 2004; Efros et al. 2003; Zelnik-Manor and Irani 2006). Then some statistics are computed on the reduced data and pattern recognition techniques such as principal component analysis (BenAbdelkader et al. 2004), bilinear models (Lee and Elgammal 2004), Hidden Markov Models (He and Debrunner 2000; Kale et al. 2004; Wilson and Bobick 1999; Oliver et al. 2000), K-Nearest Neighbor classification (Little and Boyd 1998) or Support Vector Machines (Lee and Grimson 2002) are used to solve the classification problem.

We propose modeling the dynamics of human gaits with hybrid linear models. As opposed to standard approaches using discrete models such as Hidden Markov Models (HMMs) and their variants (Oliver et al. 2000; Wilson and Bobick 1999), hybrid models capture both the *discrete* and *continuous* character of human motion and can be used for both *synthesis* (Bissacco 2005) and *recognition*.

Inference of the state and model parameters for a switching linear model is, in general, NP complete (Tugnait 1982). While several approximations exist (e.g. Pavlovic and Rehg 2000; Oh et al. 2005; Agarwal and Triggs 2004; North et al. 2000), there is no optimal algorithm of reasonable complexity for the model orders that we need to consider. Therefore, we concentrate on a specific class of models, that is switching autoregressive (AR) ones. These are a subclass of switching linear systems that is particularly attractive since, for each model, the optimal estimator can be written as a closed-form function of the data (Ljung 1987). For hybrid-AR models, recent algebraic approaches to filtering and identification (Ma and Vidal 2005) have shown promising results; however, they do not provide probabilistic information on the estimates and therefore are not suited to our purposes. We will derive our own identification algorithm in Sect. 2.2, and this is our first contribution.

Our second challenge is to define a distance in the space of hybrid-AR models. Common approaches to model-based motion recognition (Bregler 1997; Wilson and Bobick 1999; North et al. 2000) perform classification by comparing the likelihoods of sequences given learned models. Such approaches present a number of drawbacks: Long sequences yield peaked likelihoods and weak generalization performance, there is no principled way to learn and compare models representing motion classes, and it is not possible to include learned priors on the model parameters. We propose to overcome these limitations by endowing a metric in the space of models. To the best of our knowledge, this has only been done once before Del Vecchio et al. (2003) for the case where the models are represented by deterministic unknown parameters, rather than having a distribution of them. We show that the simple extension of Del Vecchio et al. (2003) to a stochastic model yields non-sensical distances that either are non-zero when the two models are identical (see (6)), or that can be infinite for models that are arbitrarily close in the deterministic sense (see (7)). The notion of discrepancy we propose is principled in the sense that, as we show, it can be written as the Euclidean distance between optimal estimators.

The main goal in this paper is to show that *the distance between hybrid models is more discriminative than the distance between linear models that was previously used to classify gaits based on their dynamics.* While this may not be surprising at first, since hybrid-AR models are a superclass of linear models, and therefore they naturally have more modeling power, note that discriminative power usually decreases with model complexity, since we can have orbits of model parameters that yield the same output statistics. This is not the case in our model, and we show that it sharply classifies gait data where linear models yield total confusion.

2 Modeling Human Dynamics for Classification

In this section we will describe the models used to describe human gaits, we will derive system identification (learning) algorithms to estimate their parameters, and we will introduce a distance between such models, that is to be used for classification in the simplest possible form, that is using a nearest-neighbor criterion. Obviously one could employ more sophisticated classifiers, but our goal here is to introduce a distance between dynamical models, so to best evaluate its properties we keep the classifier trivial. Obviously, classification results can only be improved by using more sophisticated decision rules.

We first go through the process for the simple linear autoregressive models. This for two reasons: First, because they are simple and provide some intuition into the algorithms. Second, because the results derived there are used as building blocks for the extension of the algorithms to the case of hybrid models.

We will adopt the following notation throughout. For a matrix A, $A^\top$ denotes its transpose and $A(i,j)$ is the element of A located at row i and column j. A sequence $(y_1, y_2, \ldots, y_T)$ is indicated with the superscript notation y^T. Given a random vector x, $p(x)$ denotes its probability density function (or mass function if x is discrete) and $E[f(x)]$ denotes the expectation of $f(x)$ taken with respect to the distribution of x.

$\mathcal{N}(\mu, \Sigma)$ denotes a random vector with Gaussian joint distribution of mean μ and covariance Σ; $G(x; \mu, \Sigma)$ is the corresponding probability density evaluated at x, while $\mathcal{U}\{1, \ldots, T\}$ indicates a discrete random variable distributed

uniformly between 1 and T. We will use I_p to denote the identity matrix of dimension $p \times p$. R_i will represent the noise covariance matrices and P_i the parameter covariance matrices.

2.1 Autoregressive Models

In this section we derive a simple learning algorithm for linear autoregressive models, and introduce a distance between models. We show that the most obvious choice of distances, the Euclidean distance between parameters, or Kullback-Leibler's divergence, lead to non-sensical classification, and hence introduce our distance.

Consider a Gaussian linear time-invariant autoregressive (AR) model of order n:

$$y_t = \sum_{i=1}^{n} A_i y_{t-i} + e_t \quad y_t \in \mathbb{R}^p,\ e_t \sim \mathcal{N}(0, R) \tag{1}$$

The equation can be rewritten in normal form:

$$\begin{aligned} y_t &= \varphi_t \theta + e_t \\ \varphi_t &= \begin{bmatrix} y_{t-1} \otimes I_p & y_{t-2} \otimes I_p & \cdots & y_{t-n} \otimes I_p \end{bmatrix} \\ \theta^\top &= \begin{bmatrix} \theta_1^\top & \theta_2^\top & \cdots & \theta_p^\top \end{bmatrix} \\ \theta_i^\top &= \begin{bmatrix} A_1(i,1) & \cdots & A_1(i,p) & \cdots \\ & A_n(i,1) & \cdots & A_n(i,p) \end{bmatrix} \end{aligned} \tag{2}$$

where $\otimes$ denotes the Kronecker tensor product and I_p is the identity matrix of dimension p.

Parameter Estimation

Assuming a Gaussian prior on the parameter $\theta \sim \mathcal{N}(\theta_0, P_0)$ and given a sequence of observations $y^T = (y_1, y_2, \ldots, y_T)$, the posterior distribution of the parameter θ is (Ljung 1987):

$$p(\theta | y^T, \theta_0, P_0, R) = G(\theta; \hat{\theta}, \hat{P}) \tag{3}$$

where:

$$\begin{aligned} \hat{\theta} &= \hat{P} \left(P_0^{-1} \theta_0 + \sum_{t=1}^{T} \varphi_t R^{-1} y_t \right), \\ \hat{P} &= \left(P_0^{-1} + \sum_{t=1}^{T} \varphi_t R^{-1} \varphi_t^T \right)^{-1} \end{aligned} \tag{4}$$

and $G(\theta; \hat{\theta}, \hat{P})$ is the Gaussian density with mean $\hat{\theta}$ and variance $\hat{P}$ evaluated at θ:

$$\begin{aligned} G(\theta; \hat{\theta}, \hat{P}) = (2\pi)^{-\frac{d}{2}} \det(\hat{P})^{-\frac{1}{2}} \\ \times \exp\left(-\frac{1}{2} (\theta - \hat{\theta})^T \hat{P}^{-1} (\theta - \hat{\theta}) \right) \end{aligned} \tag{5}$$

For an intuitive understanding of these expressions consider the simple case of scalar measurements $y \in \mathbb{R}$. The equation of $\hat{P}$ reduces to: $\hat{P} = \left(P_0 + \frac{\sum_{t=1}^{T} y_t^2}{R} \right)^{-1} = \left(P_0 + (T-1) \frac{\Sigma_y}{R} \right)^{-1}$, where Σ_y is the sample variance of the measurements. The variance $\hat{P}$ is a measure of the uncertainty in the estimated parameters. As we could expect, it decreases as the length T of the observation sequence and the signal-to-noise ratio $\frac{\Sigma_y}{R}$ increase. In the limit $T \to \infty$, the variance $\hat{P}$ becomes zero and the estimate $\hat{\theta}$ is the true value of the parameters.

Model Distance (AR)

We use the posterior distributions $p(\theta | y^T)$ on the parameters to define a distance between models. As a first attempt we consider the expectation of the Euclidean distance between the parameters $\theta_1 | y_1^T \sim \mathcal{N}(\hat{\theta}_1, \hat{P}_1), \theta_2 | y_2^T \sim \mathcal{N}(\hat{\theta}_2, \hat{P}_2)$:

$$\begin{aligned} d_e(\theta_1, \theta_2)^2 &= E[(\theta_1 - \theta_2)^\top (\theta_1 - \theta_2)] \\ &= (\hat{\theta}_1 - \hat{\theta}_2)^\top (\hat{\theta}_1 - \hat{\theta}_2) + \mathrm{Trace}(\hat{P}_1 + \hat{P}_2) \end{aligned} \tag{6}$$

Unfortunately, this is not a viable distance between models; indeed, it is not even a distance, in the sense that $d_e(\theta_1, \theta_1) \neq 0$, violating one of the conditions that define a distance function. A second attempt is to consider as a discrepancy function the symmetric Kullback-Leibler divergence (K-L) between the two distributions:

$$KL(p_1 \| p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} + p_2(x) \log \frac{p_2(x)}{p_1(x)} dx \tag{7}$$

which for our Gaussians becomes:

$$\begin{aligned} KL(\theta_1 \| \theta_2) = \frac{1}{2} \mathrm{Trace} \left(\hat{P}_2^{-1} \hat{P}_1 + \hat{P}_1^{-1} \hat{P}_2 - 2I \right) \\ + (\hat{\theta}_1 - \hat{\theta}_2)^\top \left(\hat{\Sigma}_1^{-1} + \Sigma_2^{-1} \right) (\hat{\theta}_1 + \hat{\theta}_2). \end{aligned} \tag{8}$$

This is also not a distance, as it does not satisfy the triangular inequality. Furthermore, as the variances $\hat{\Sigma}_1, \hat{\Sigma}_2$ go to zero (i.e. the confidence on the parameter estimates increases), the divergence goes to infinity. K-L is a measure of the extent to which two probability distributions agree. If the two distributions have no common support the K-L distance is infinite independently of how far the distributions are (see the example in Fig. 2). Such a condition is met for example when we have good estimates from sequences generated by models with different underlying parameters.

We can overcome these problems by using a distance between probabilities distributions known with several names, as the Wasserstein, Mallows, Ornestein, or rho-bar distance (Bickel and Freedman 1981). Using the L_2 norm as base

distance, it is defined between two densities P and Q as:

$$d_W(P,Q)^2 = \inf_F\{E_F[(X-Y)^\top(X-Y)]: (X,Y)\sim F, X\sim P, Y\sim Q\} \quad (9)$$

where the infimum is taken over all the joint densities F 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 Gaussian distributions d_W can be computed analytically as in (Dowson and Landau 1982):

$$d_W(\mathcal{N}(\hat{\theta}_1,P_1),\mathcal{N}(\hat{\theta}_2,P_2))^2 = (\hat{\theta}_1-\hat{\theta}_2)^\top(\hat{\theta}_1-\hat{\theta}_2)+\mathrm{Tr}(P_1+P_2-2(P_1P_2)^{\frac{1}{2}}) \quad (10)$$

This distance has some desirable properties. First it is a proper distance, in particular it satisfies the triangular inequality. This guarantees that if the estimated densities $\hat{P}$, $\hat{Q}$ are good (i.e. $d(P,\hat{P})$ and $d(Q,\hat{Q})$ are small), also the estimated distance $d(\hat{P},\hat{Q})$ is close to the true distance $d(P,Q)$: $|d(P,Q)-d(\hat{P},\hat{Q})|\leq|d(P,\hat{P})|+|d(Q,\hat{Q})|$. Second, it is equal to the Euclidean distance in the case of deterministic distributions $P_1=P_2=0$.

For discrete distributions, the Wasserstein distance is equivalent to the Earth's movers distance (Rubner et al. 1998), a distance commonly used for measuring texture and color similarities.

In the more general case of a mixture of Gaussian distributions, no close form solution is available and an approximation must be used, as we will show in the next section.

2.2 Hybrid Autoregressive Models

In order to model contact forces in human motion we follow the approach of Bissacco (2005) in using hybrid models where the switches correspond to ground contacts. However, unlike Bissacco (2005), we intend to use such models for classification, and therefore we introduce a different switching autoregressive model. This has some similarity with the Autoregressive HMM proposed in Juang and Rabiner (1985), although for each autoregressive model we consider the distribution of the observations y_t for finite length sequences instead of using the asymptotic distribution of $y_t, t\to\infty$.

Consider a discrete Markov chain with m states, transition matrix M and prior probabilities $\pi^m=[\pi_1,\ldots,\pi_m]$. To each state q we associate an AR model with noise covariance R_q and parameter θ_q with prior distribution $\theta_q\sim\mathcal{N}(\theta_{0,q},P_{0,q})$. The equations of the system are:

$$\begin{aligned} &y_t=\varphi_t\theta_{q_t}+e_{q_t}, \quad e_{q_t}\in\mathcal{N}(0,R_{q_t}) \\ &p(q_t|q_{t-1})=M(q_t,q_{t-1}), \quad p(q_1)=\pi_{q_1} \end{aligned} \quad (11)$$

A graphical representation of this model is shown in Fig. 1. As we can see from the figure, the AR parameters $\theta^m=(\theta_1,\ldots,\theta_m)$ are time-invariant random vectors, and the observed outputs y_t induce a distribution on hidden states q_t and model parameters θ^m. The motivation for this model is that we assume m underlying autoregressive models, whose parameters θ_i are random but fixed, and the transitions between models are determined by the hidden states q_t. Related models have been proposed in the adaptive filtering literature, where for each segmentation of the output sequence a different linear regression model is assumed in each segment and the posterior of the segmentation is computed by marginalizing the hidden parameters (Gustafsson 2000). Our model is more complex in that we assume a finite number m of autoregressive systems, and the transitions among these are governed by a Markov chain. Thus, as opposed to Gustafsson (2000), observation segments are no longer statistically independent given the segmentation, which makes the inference problem harder.

In other hybrid AR systems proposed in the literature (Del Vecchio et al. 2003; North et al. 2000), the parameters θ are modeled as unknown deterministic values. A learning algorithm is derived to compute the maximum likelihood estimate $\theta^{ML}=\arg\max_\theta p(y^T|\theta)$ given an observation sequence y^T. Unfortunately, this method does not provide a natural way to compare the parameters of two models θ_1,θ_2, and a common solution (Del Vecchio et al. 2003) is to use the Euclidean distance between the parameters, $\|\theta_1-\theta_2\|$. Our approach is different in the sense that we treat θ as a random vector with given prior distribution $p(\theta)$ and compute the posterior given the observations $p(\theta|y^T)$. This allows us to consider multiple model hypotheses by inferring (multimodal) posteriors on the model parameters and comparing models by using distances between these probability distributions.

We can relate the two approaches by considering the case of flat (uninformative, possibly improper) prior $p(\theta)\simeq$ const. Then the posterior $p(\theta|y^T)$ is proportional to the likelihood $p(y^T|\theta)$, and the maximum likelihood estimate is also the maximum a posteriori $\theta^{ML}=\hat{\theta}$. The distance $d^{ML}=\|\hat{\theta}_1-\hat{\theta}_2\|$ measures how far the principal modes of the posterior distributions $p(\theta_1|y^T)$ and $p(\theta_2|y^T)$ are. In the case of hybrid models this solution is suboptimal since the posteriors $p(\theta_i|y^T)$ are typically multimodal mixtures, as we can see in Fig. 6, while the distance d^{ML} takes into account only one parameter hypothesis.

Once the reader accepts the use of hybrid models for human gaits, we are left with the problem of learning the parameters and filtering its state, and then extend the definition of distance, so we can perform classification tasks. We discuss these problems in the next two subsections.

2.2.1 Parameter Estimation

Given an observation sequence $y^T = (y_1, \ldots, y_T)$ generated by the hybrid model (11), we want to estimate the posterior distribution of the autoregressive parameter θ randomly sampled from the sequence $(\theta_{q_1}, \ldots, \theta_{q_T})$:

$$\theta \triangleq \theta_{q_\tau}, \quad \tau \sim \mathcal{U}\{1, \ldots, T\} \tag{12}$$

We have:

$$\begin{aligned} p(\theta|y^T, \Lambda) &= \sum_{q^T} p(\theta_{q_\tau}|q^T, y^T, \Lambda) p(q^T|y^T, \Lambda) \\ &= \sum_{q^T} \sum_{i=1}^{m} p(\theta_i|q^T, y^T, \Lambda) p(q_\tau = i|q^T, y^T, \Lambda) \\ &\quad \times p(q^T|y^T, \Lambda) \\ &= \sum_{q^T} \sum_{i=1}^{m} p(\theta_i|q^T, y^T, \Lambda) p(q_\tau = i|q^T) \\ &\quad \times p(q^T|y^T, \Lambda) \end{aligned} \tag{13}$$

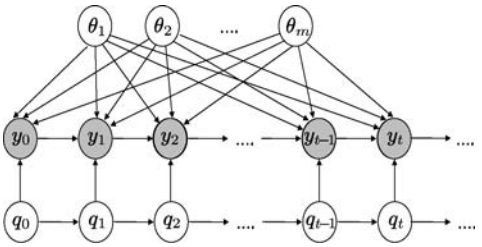


Fig. 1 Dynamic Bayesian network representing our proposed hybrid autoregressive model. Nodes are random vectors (observed nodes are shaded) and edges are conditional dependence relations. The presence of multiple loops in the graph makes exact inference a computationally intractable problem

where $\Lambda = \{\theta_0^m, P_0^m, R^m, M, \pi\}$ are the model parameters, with $\theta_0^m = (\theta_{0,1}, \ldots, \theta_{0,m})$, $P_0^m = (P_{0,1}, \ldots, P_{0,m})$, $R^m = (R_1, \ldots, R_m)$. Similarly to (3), we have that $p(\theta_i|q_j^T, y^T, \Lambda) = G(\theta_i; \hat{\theta}_i, \hat{P}_i)$ are Gaussian, $p(q_\tau = i|q^T)$ is the relative frequency of the state i in the sequence q^T, and $p(q^T|y^T, \Lambda)$ is the posterior of the hidden states given the observations, that can be computed in closed form as we will show in the next section. Unfortunately, marginalizing the hidden states $q^T = (q_1, \ldots, q_T)$ is intractable because it would require evaluating an exponential number of hypotheses.

Let us point out that existing techniques for learning hybrid models similar to (11), such as (North et al. 2000; Ghahramani and Hinton 1998; Pavlovic and Rehg 2000; Oh et al. 2005), treat the system matrices as parameters and not as latent variables, therefore such approaches cannot be applied to the problem of computing parameter posteriors.

A first approach to inference would be to apply Gibbs sampling to obtain sequences of hidden states q^T and model parameters θ^m distributed according the posterior. However, we have observed that for this model the parameters typically have highly peaked multimodal distributions, trapping the Gibbs sampler in local modes and thus requiring large number of samples to obtain good approximations.

On the other hand, the graphical model in Fig. 1 shows that each parameter θ_i is statistically dependent on all the observations y_t, thus preventing the application of standard algorithms such as loopy belief propagation.

We could apply variational inference techniques in order to obtain an approximate model with a smaller number of dependencies for which the inference problem would be easier. Typically these methods work by approximating the posterior of the hidden variables q^T given observations y^T

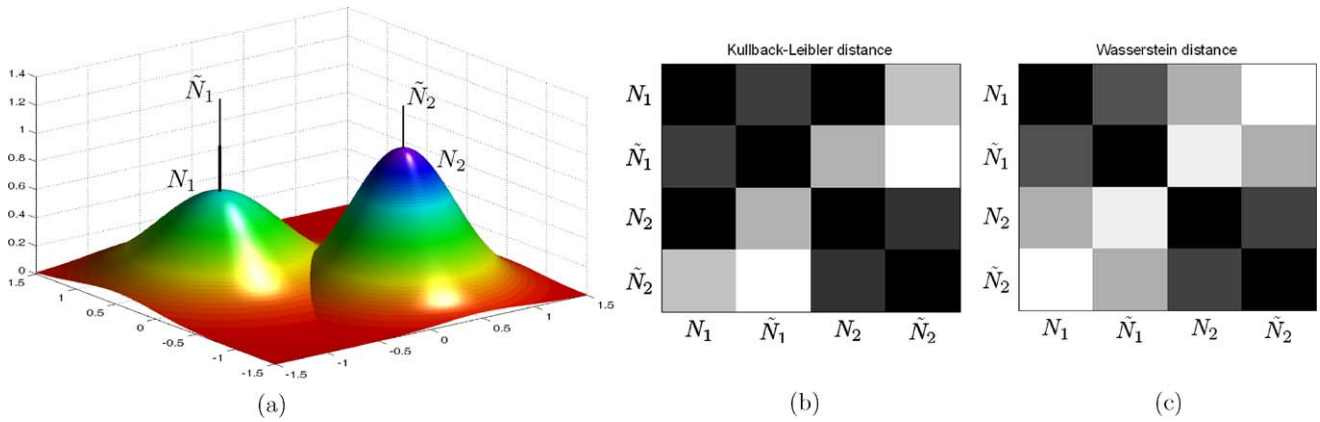


Fig. 2 Comparison between Wasserstein distance and Kullback-Leibler divergence for classifying parameter distributions. (**a**) Simple case of 4 Gaussians, where $N_1, \tilde{N}_1$ have the same mean but $\tilde{N}_1$ has smaller variance, similarly for N_2 and $\tilde{N}_2$. (**b**) Confusion matrix showing the pairwise Kullback-Leibler divergence between these distributions. We can see that Kullback-Leibler has the undesirable effect of classifying N_2 closer to N_1 than to $\tilde{N}_1$, whereas N_1 and $\tilde{N}_1$ represent the same estimate with different degree of confidence. (**c**) Confusion matrix from the Wasserstein distance: Here $N_1, \tilde{N}_1$ and $N_2, \tilde{N}_2$ are correctly grouped together, and the distance between $\tilde{N}_1$ and $\tilde{N}_2$ equals the distance between their means as the variance goes to zero (as opposed to becoming infinite as the Kullback-Leibler does)

and parameters θ^m. However, notice that by doing so there is no simple way to break the dependencies between outputs and parameters, therefore we would not remove the main source of complexity in the model.

Our solution is to approximate the posterior using a bank of K filters, where each filter is tuned on a segmentation hypothesis q_j^T. At each time t we generate a new hypothesis q_t^T by imposing a jump to the most likely sequence and discarding the less likely ones. In formulas, this corresponds to the following approximation:

$$p(\theta|y^T,\Lambda) \simeq \frac{1}{C}\sum_{j=1}^{K}\sum_{i=1}^{m} p(\theta_i|q_j^T,y^T,\Lambda)p(q_\tau=i|q_j^T) \times p(q_j^T|y^T,\Lambda) \tag{14}$$

where $C=\sum_{j=1}^{K} p(q_j^T|y^T,\Lambda)$ and q_j^T are the filter hypotheses. This approximation is a mixture of a constant number Km of Gaussians. In practice, we have duplicate hypotheses (due to permutations of the states) and hypotheses with low posterior, so the effective number of components can be rather smaller (Fig. 6).

Hidden State Filtering

In order to obtain a good approximation of the posterior (14) we need to estimate the K most probable hidden state sequences $q_1^T,\ldots,q_K^T$ given the measurements y^T:

$$\hat{q}_1^T,\ldots,\hat{q}_K^T = \arg\max_{q_1^T,\ldots,q_K^T}\sum_{i=1}^{K} p(q_i^T|y^T,\Lambda) \tag{15}$$

Notice that, since in our model the autoregressive parameters θ_i conditioned on the measurements y^T are Gaussian, it is possible to marginalize them in the computation of the hidden state posterior (15). This is a remarkable advantage compared to approaches with deterministic parameters such as North et al. (2000) where Expectation Maximization or other iterative minimization techniques become necessary, because the resulting segmentation no longer depends on the initial guess. Here only a prior is needed, and using an uninformative one such as a Gaussian with high variance allows for a non-iterative, unbiased estimation of the switching times.

We derive a recursive expression for the likelihood $p(y^T|q^T,\Lambda)$:

$$p(y^t|q^t,\Lambda) = p(y_t|q^t,y^{t-1},\Lambda)p(y^{t-1}|q^{t-1},\Lambda) \tag{16}$$

which yields (see Gustafsson 2000):

$$p(y_t|q^t,y^{t-1},\Lambda) = G(y_t;\varphi_t^\top\hat{\theta}_{q_t,t-1},\varphi_t^\top\hat{P}_{q_t,t-1}\varphi_t+R_{q_t}) \tag{17}$$

where $\hat{\theta}_{i,t}$, $\hat{P}_{i,t}$ are the estimates at time t of the parameters θ_i associated to state i (compare to (4)):

$$\hat{\theta}_{i,t} = \hat{P}_{i,t}\left(P_{0,i}^{-1}\theta_{0,i} + \sum_{j|q_j=i,j\le t}\varphi_j R_i^{-1} y_j\right) \tag{18}$$

$$\hat{P}_{i,t} = \left(P_{0,i}^{-1} + \sum_{j|q_j=i,j\le t}\varphi_j R_i^{-1}\varphi_j^\top\right)^{-1} \tag{19}$$

which can be rewritten in recursive form as:

$$\hat{\theta}_{i,t} = \hat{\theta}_{i,t-1} + \hat{P}_{i,t-1}\varphi_t(\varphi_t^\top\hat{P}_{i,t-1}\varphi_t+R_i)^{-1} \times (y_t-\varphi_t^\top\hat{\theta}_{t-1}) \tag{20}$$

$$\hat{P}_{i,t} = \hat{P}_{i,t-1} - \hat{P}_{i,t-1}\varphi_t \times\left(\varphi_t^\top\hat{P}_{i,t-1}\varphi_t+R_i\right)^{-1}\varphi_t^\top\hat{P}_{i,t-1} \tag{21}$$

From (16) we obtain a recursive equation for the posterior up to time t:

$$\begin{aligned} p(q^t|y^t,\Lambda) &= \frac{1}{K_t}p(y^t|q^t,\Lambda)p(q^t|\Lambda) \\ &= \frac{K_{t-1}}{K_t}p(y_t|q^t,y^{t-1},\Lambda)p(q_t|q_{t-1},\Lambda) \\ &\quad\times p(q^{t-1}|y^{t-1},\Lambda) \end{aligned}$$

where $K_t = p(y^t|\Lambda)$ is a constant independent of q^t, $p(q_t|q_{t-1},\Lambda)=M(q_{t-1},q_t)$, $t>1$ and $p(q_1|q_0,\Lambda)=\pi_{q_1}$. Substituting $p(y_t|q^t,y^{t-1},\Lambda)\sim\mathcal{N}(\varphi_t^\top\hat{\theta}_{q_t,t-1},\varphi_t^\top\cdot\hat{P}_{q_t,t-1}\varphi_t+R_{q_t})$, and taking the logarithms, we have:

$$\begin{aligned} \log p(q^t|y^t,\Lambda) &= C+\log p(q^{t-1}|y^{t-1},\Lambda) \\ &\quad+\log M(q_{t-1},q_t)-\frac{1}{2}\log\det\Gamma \\ &\quad-\frac{1}{2}\left(y_t-\varphi_t^\top\hat{\theta}_{q_t,t-1}\right)^\top \\ &\quad\times\Gamma^{-1}\left(y_t-\varphi_t^\top\hat{\theta}_{q_t,t-1}\right) \end{aligned} \tag{22}$$

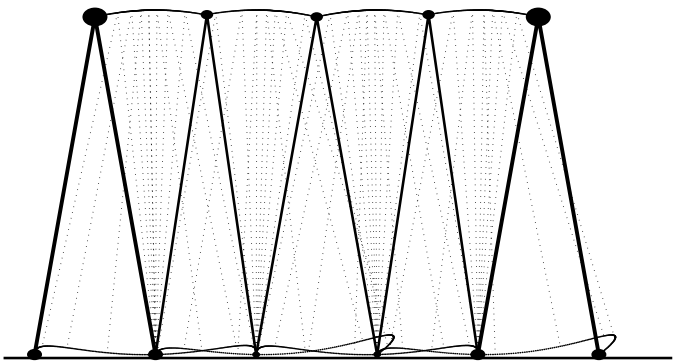
where C is a constant and $\Gamma=(\varphi_t^\top\hat{P}_{q_t,t-1}\varphi_t+R_{q_t})$.

To find an approximate solution to the most probable state sequences problem (15), we use a bank of K filters, each matched to a hidden state sequence hypothesis q_i^T, where the posterior $p(q^T|y^T,\Lambda)$ is computed recursively with (22). We use Algorithm 1.

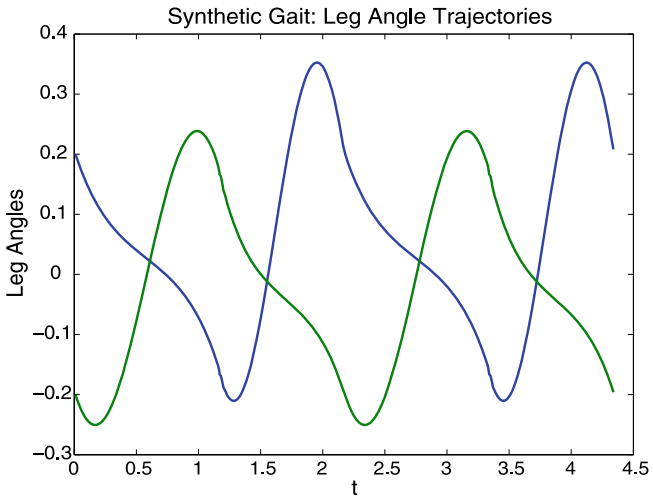
The number of filters K determines the quality of the estimates. For $K\ge T$, this algorithm is guaranteed to find the optimal state sequences for switching regression models where data segments are independent (Gustafsson 2000). However, in our hybrid models, different segments can be

Algorithm 1 Hybrid autoregressive model hidden state filtering using marginalized posteriors

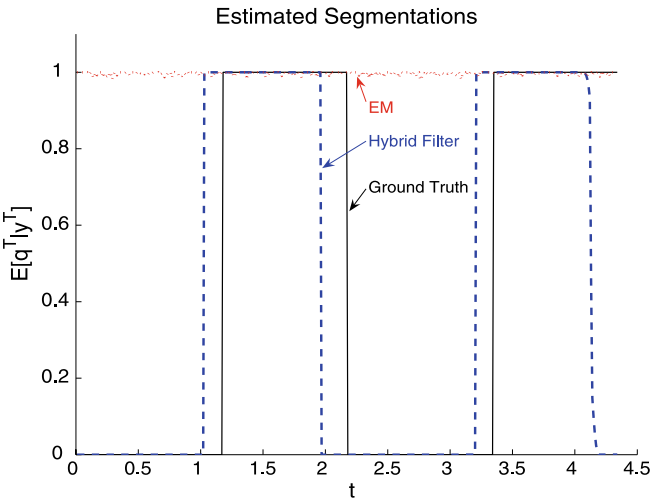
1: Initialize states so that there is at least one sample $q_i^1 = i$ for each $i \in \{1, \ldots, m\}$.
2: **for** $t = 2$ to T **do**
3: For each hypothesis q_i^t, compute the posterior log-likelihood $\log p(q_i^t | y^t)$ using (22).
4: Extend the hypotheses q_j^t, $j = 1, \ldots, K$ to $t+1$ by assuming no switch: $q_j^{t+1} = (q_j^t, q_{j,t})$.
5: Let the most probable sequence q_o^t split at time $t+1$, i.e. generate $m-1$ new hypotheses q_{K+i}^{t+1} such that $\{q_{K+i,t+1}\} = \{1, \ldots, m\} \backslash \{q_{o,t}\}$.
6: Cut off the $m-1$ least probable sequences, so that only K are left.
7: **end for**



(a)



(b)



(c)

Fig. 3 Learning hybrid model segmentations from synthetic walk data. (**a**) Simple 2D passive walker model (Garcia et al. 1998) used to generate the synthetic motion. An asymmetric walk is obtained by setting unit mass at the hip, 0.2 at the left and 0 at the right foot. (**b**) Two gait cycles represented as trajectories of the angles of the legs with respect to the vertical. (**c**) Expected posteriors on the hidden states computed from the segmentation hypothesis generated by our filtering approach (*solid*) and the EM algorithm after 100 iterations (*dotted*), with switching ground truth. The plot shows how the EM approach fails to segment the dynamics of the gait motion, whereas our hybrid filter finds the switching sequence. There is a slight offset between the switches in the ground truth and ones obtained from the hidden state estimates. Because the identification is performed in a non-causal fashion, the offset can be either positive (delay) or negative. This error may be due to the linear approximation, given that the nonlinear dynamics of each motion segment is represented by a first order autoregressive model

generated by the same autoregressive model, and therefore the estimated sequences $\hat{q}_i^T$ will only be an approximate solution to (15). In order to improve performance it is useful to assume a minimum segment length l and allow splitting and cut off only for sequences that did not switch in the last l steps.

In the experiments we compare our filtering approach to learning hybrid autoregressive models with deterministic parameters via EM and particle-filtering as proposed in North et al. (2000). By applying both techniques to segment synthetic data of a simple passive walker, in Fig. 3 we show that our hybrid filters can successfully estimate the switching times whereas the EM approach fails to segment the hybrid dynamics.

Distance Between Hybrid AR Models

We obtain a discrepancy measure between models by extending to hybrid models the distance (10) between posteriors of autoregressive parameters. For each motion sequence y_k^T, let the posterior distribution of the hybrid parameter θ_k be approximated as:

$$p(\theta_k|y_k^T, \Lambda) \simeq \sum_{i=1}^{n_k} \alpha_{k,i} G(\theta_k; \hat{\theta}_{k,i}, \hat{P}_{k,i}) \tag{23}$$

The Wasserstein distance between general mixtures of Gaussians cannot be computed in closed form. Following Greenspan et al. (2004), we approximate $d_W(\theta_1, \theta_2)$ by solving a maximum flow problem. We have:

$$d_W(\theta_1, \theta_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f_{i,j} d_W(\mathcal{N}(\hat{\theta}_{1,i}, \hat{P}_{1,i}), \mathcal{N}(\hat{\theta}_{2,j}, \hat{P}_{2,j})) \tag{24}$$

where the Wasserstein distance between normal distributions is given in (10) and $f_{i,j} \geq 0$ is the optimal admissible flow that minimizes (24) while satisfying the constraints:

$$\sum_{j=1}^{n_2} f_{i,j} = \alpha_{1,i}, \qquad \sum_{i=1}^{n_1} f_{i,j} = \alpha_{2,j}$$

In the next section we will use this distance to compare hybrid dynamical models learned from human motion data.

3 Experiments

Our goal in this research is to recognize human motion based on dynamic signatures. We believe that the temporal evolution of a representation abstracted from video carries a significant amount of discriminative power: Johansson's moving-dot displays (Johansson 1973) can allow one to infer whether the person is young, old, happy, sad, even man or woman. In particular, we use a hybrid dynamical model because we have determined that the contact dynamics, which is an exogenous event independent of the individual and his/her gait, is a dominant dynamic event that must be factored out of the classification and recognition process. However, our framework applies to the recognition of dynamic events in general. In particular, even within human motion, our framework applies to different representations, from the trajectories of moving intensity blobs, to the joint angles estimated from a video-based tracking system, to the position of retro-refractive markers in motion capture. Let us also point out that hybrid generative models as in Bissacco (2005) can be effectively used for synthesis, while common discrete models (e.g. HMMs) cannot reproduce the original data except at a very coarse granularity.

In order to perform a fair comparison with competing approaches, we must factor out the effects of photometric and geometric factors in the image formation process. For this reason we concentrate on motion capture data. Comparison with approaches based on silhouette extraction, block-correlation or other appearance-based approach that does not use a dynamical model would not shed light on the virtues or limitations of our approach.

To the best of our knowledge, the only other class of dynamical models used for motion analysis is linear. While on one hand linear models can capture the second-order statistics of any stationary sequence arbitrarily well (indeed, for a large-enough state one can approximate the actual realizations to an arbitrary degree), we will show that—at equal model order—our approach substantially improves classification when compared to linear dynamical models. This is because most of the energy in the data occurs at contact events, and therefore most of the modeling power goes to represent such phenomena, that are nuisance factors irrelevant to classification.

We performed two sets of experiments. The first is on synthetic data, where the goal is comparing our learning approach based on hidden state filtering to standard maximum likelihood techniques based on Expectation Maximization. The second set of experiments aims at comparing the performance in gait discrimination using the proposed distance between hybrid systems with respect to previous approaches based on distances between single linear dynamical systems.

3.1 Hidden State Filtering

In this experiment we test the performance of the hidden state estimation of our filtering approach compared to standard parameter estimation by EM and particle filtering, with a method similar to the one proposed in North et al. (2000). As the reader may notice, our model differs from the one of North et al. (2000) in that the y_t is observed instead of being a hidden variable, thus the forward-backward filtering proposed in North et al. (2000) is not needed. Even despite this simplification, the dependencies between observations $y_1, \ldots, y_T$ do not allow to derive a backward recursion for the hidden state likelihood (conditioning on q_t does

Subject	1	2	3	4	5	6	7	8	9	10	11	Total
Walk	14	14	7	14	14	2	5	14	7	14	9	114
Run	14	8	9	14	8	-	7	3	-	8	8	79
Limp	4	2	4	5	5	-	5	5	-	5	5	40
Total	32	24	20	33	27	2	17	22	7	27	22	233

Fig. 4 List of motion capture data sequences in the gait dataset. For each subject (*first column*), number of walking, running and limping sequences collected

not make y_t and y_{t+1} independent, see Fig. 1), therefore we cannot solve the Expectation step in polynomial time and a sampling scheme is still required.

In order to have controlled conditions with ground truth, we use synthetic gait data generated by the simple passive walker of Garcia et al. (1998). It is a 2D two-link model, with rigid massless legs hinged at the hip and point masses at the feet, that walks down a slope under the effects of gravity. We set unit mass at the hip and unequal masses at the feet, respectively 0 and 0.2, so as to have a limp-like motion. The data consists of the trajectories of the leg angles in two walking cycles, as shown in Fig. 3. We approximate such time series with a two-system (left and right step) hybrid autoregressive model of order $n = 1$.

We applied our filtering Algorithm 1 and the Expectation-Maximization algorithm with particle filtering to estimate hidden parameters θ^m and discrete states q^t. We use Algorithm 2, a simple particle filter based on the forward propagation of North et al. (2000), which provides an efficient way to sampling sets of K particles on-line for $t = 1, \ldots, T$ $\hat{q}_t(1), \ldots, \hat{q}_t(K)$ from the posterior $p(q^T|y^T, \varphi^m, R^m)$.

The complexity of the Expectation step is $O(KT)$, where K is the number of samples. The Maximization step is the same of North et al. (2000), where the expected values are used to update the estimates of the parameters θ^m, M, π (as suggested in North et al. 2000 the noise variances R^m are set by hand and not learned). In this experiment we use a hybrid model with two systems, the number of samples is equal to the sequence length $K = T$, the model parameters θ^m are initialized randomly and the Markov chain parameters are so that all states have equal probability and average segment length L: $M(i, j) = \frac{1}{(L+1)(m-1)}$ $i \neq j$, $w_i = \frac{1}{m}$, here $L = 50$. The noise variance is fixed and set to $R_q = Ir$ with r a random variable uniformly distributed in $[0, 10^{-3}]$.

In the hybrid filter approach of Algorithm 1, we use the same parameters, and set the number of filters K equal the sequence length T. The time complexity of the algorithm is $O(TK)$.

Figure 3 shows the segmentation results on the synthetic walker data. We compare the expected posteriors on the hidden states estimated by the two approaches, computed from the sample segmentations $\hat{q}_i^T$ as $E[q^T|y^T, \Lambda] \simeq \sum_{i=1}^K p(\hat{q}_i^T)\hat{q}_i^T$. We can see how the EM approach fails to separate the different dynamics of the two phases of the gait cycle, attributing the largest part of the sequence to a single system. This is an example of how such gradient approaches can get stuck in local minima if the initial guess on the parameters is not close to the true value. On the other hand, our filtering approach avoids any iterative minimization scheme by marginalizing out the system parameters, which in our model are hidden variables. As we see in Fig. 3, it successfully finds the correct segmentation, and the computational cost amounts to a single iteration of the particle-based EM method.

3.2 Gait Classification

In the second set of experiments, the data is given as a set of joint angle trajectories on a skeletal model of the human body. These angles may be obtained from a video-based full body tracker or from a motion capture system. We opted for the latter for ease of collection and ground-truth testing. We used a 6-camera infrared motion capture system running at 60 Hz, with 20 retro-reflective markers placed on the test subjects at the proximity of the body joint locations, and with that we recorded the marker trajectories during the motion. The subjects were asked to walk, run and limp on a treadmill. We collected a total of 233 sequences from 11

Algorithm 2 Expectation step in hybrid autoregressive model learning

for $k = 1$ to K **do** {Initialization}
  $w_0(k) = \frac{1}{K}$
  Sample $q_0(k) \sim \mathcal{U}\{1, \ldots, m\}$
end for
for $t = 1$ to T **do**
  for $k = 1$ to K **do** {Online Sampling}
    Resample from observation weights: $l \sim w_{t-1}$
    Predict $q_t(k)$ by sampling from $p(q_t|q_{t-1}(l))) = M(q_{t-1}(l), q_t)$
    Compute new observation weights: $w_t(k) = G(y_t - \varphi_t\theta_{q_t(k)}; 0, R_{q_t(k)})$
  end for
end for

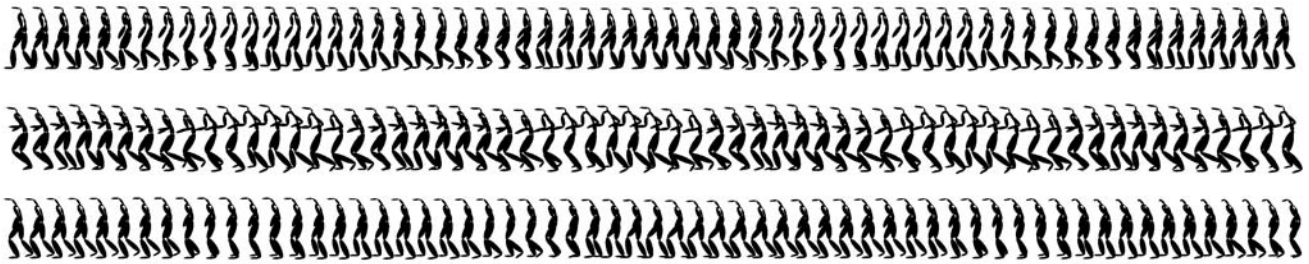


Fig. 5 Short clips (about 3 seconds) from the sequences of the gait dataset. Subject 1 walking (*top*), running (*center*) and limping (*bottom*)

subjects, see the table in Fig. 4 for details. Each sequence is sampled at 60 Hz and is about 6 second-long. Different sequences thus sample different instance of the same gait performed by different individuals.

From marker positions we estimated body skeleton model and joint angles with an approach similar to the one proposed in O'Brien et al. (2000). First, we estimated the reference frame moving with the body limb from the set of markers attached to the limb. Then the joint positions were obtained as the center of rotation of the reference frame of adjacent limbs. From joint positions, by enforcing fixed limb length, we obtained the model of the skeleton and the joint angles. Since we did not use a reference model for the skeleton, the estimated skeletons vary from person to person, affecting the joint angles estimates and making the recognition problem harder. In Fig. 5 we show some sample clips of the data sequences. From each sequence, we extracted the 24 angles corresponding to the 8 joints defining the positions of hips, femurs, tibias and feet. The angles are represented using the exponential map (Ma et al. 2003). Since the number of parameters of the AR model is p^2, where p is the dimension of the measurements, we had to reduce the dimensionality of the data. For this purpose we applied principal component analysis (PCA) to each sequence, treated as a collection of static poses. We retained the first $p = 4$ components, and used the coefficients of the joint angles projected onto the learned basis as observations y_i^T. From the low-dimensional sequence y_i^T, we learned the posterior (14) using Algorithm 1 described in Sect. 2.1. The model we propose is very general and contains a number of parameter that should be tuned to the particular class of signals under investigation. In these experiments, we used first-order autoregressive models, i.e. $n = 1$ in (1). Of course, in order to take into account acceleration, a second-order system would be more appropriate. However, in these experiments, we show that a first-order model is sufficient for discriminative purposes. We set the prior means θ_0^m to zero and the prior variances P_0^m to $p_0 I$, where p^0 is a large number, to capture the lack of prior information on the parameters. The noise variances R^m are set to the identity, so that in (18) we obtain least squares estimates. As before, we have 2 hidden states and models with equal probability. The posteriors are computed with a bank of K filters. To have optimal segmentations we would need K to be no smaller than the sequence length T, typically about 400. In practice, we noticed that reducing K to 50 does not significantly change the approximation (14). Since some of the computed segmentation hypotheses are equivalent (they are equal up to a permutation of the states), the filtering is followed by a hypothesis reduction step where we remove the duplicate hypotheses. Then we proceed to compute the posterior on the parameters (14). Of all the components of (14), typically only few have weight significantly different from zero. Therefore, we proceed by pruning all the hypotheses that have weight below a small threshold. In Fig. 6 we show the weight of the mixture components of the parameter posteriors learned from the gait sequences. We see that most of the sequences have multimodal distribution, with a number of modes limited by the number of filters K.

3.3 Hybrid Models for Dynamic Discrimination

The point of this section is to show that hybrid models have more discriminative power than simpler linear models (Bissacco et al. 2001). Our intent here is to show that discrimination between different classes (e.g. different gaits by the same individual, or different individuals walking the same gait) is made possible by a hybrid model where it was not by using a linear dynamical model.

This is, therefore, a feasibility study, not in competition with other gait or individual recognition techniques that use different (static) features. Our approach is meant to complement them, not to replace them.

In Fig. 6 we show the pairwise distance between models learned from the dataset sequences. We clearly see that the hybrid models can discriminate between gait classes. For comparison, we learned first-order autoregressive models from the same sequences and computed the Euclidean distance between the maximum likelihood parameter estimates. By using this simpler model distance we would not be able to discriminate between gaits. The confusion between limp and walk may be due to the different parameterizations of the motion, to the dimensionality reduction

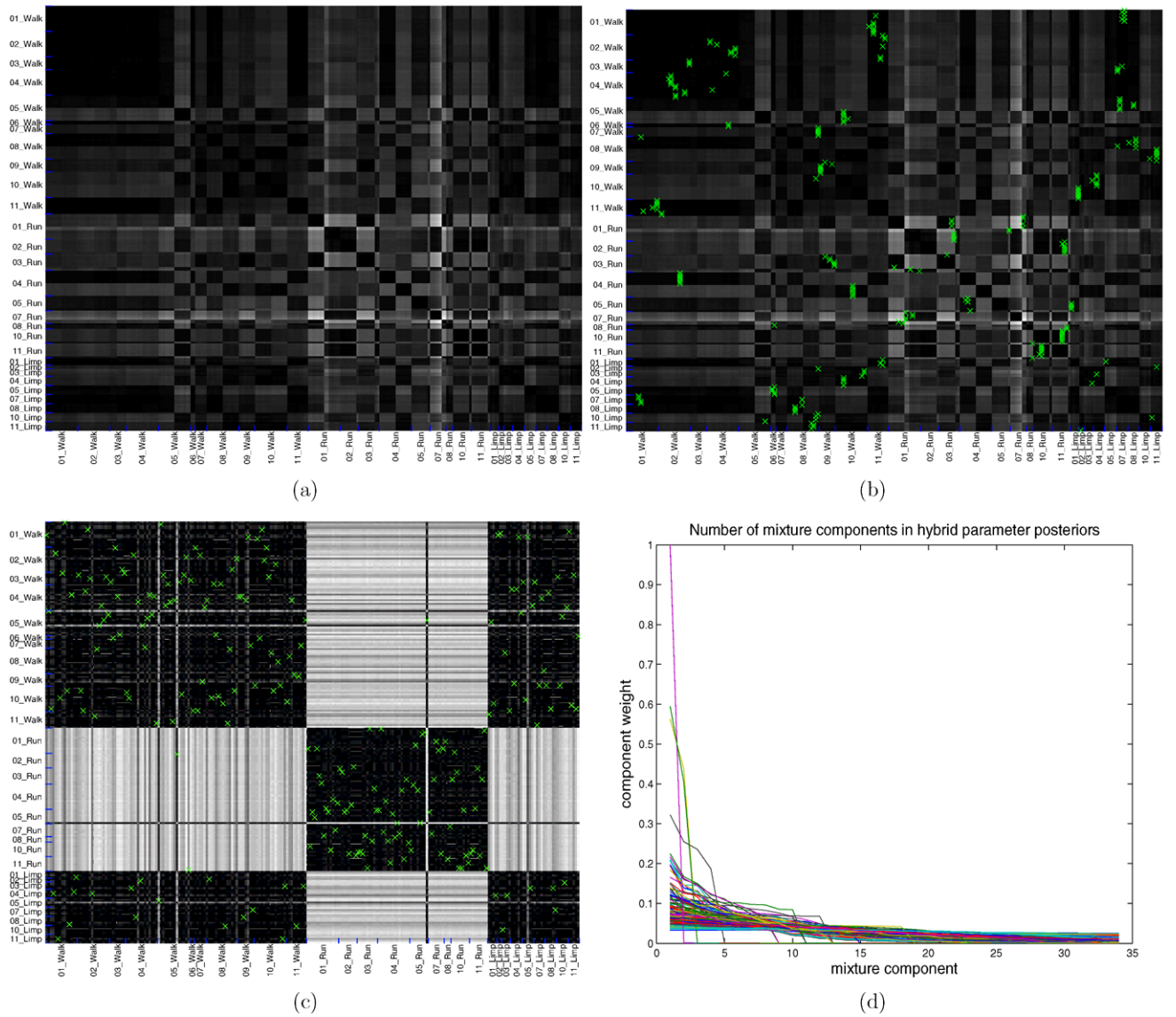


Fig. 6 (Color online) Discrepancy measure between models learned from the gait dataset; (**a**) shows the Euclidean distance between maximum likelihood estimates of autoregressive model parameters; (**b**) displays the Martin distance between ARMA models as proposed in Bissacco et al. (2001); (**c**) shows the approximated Wasserstein distance between posterior distributions of the parameters of the hybrid autoregressive models. For each row the *green cross* denotes the nearest neighbor. We can see that the simple autoregressive models are not discriminative enough to capture the character of the motion class. Following Bissacco et al. (2001), we increase the descriptive power by using state-space ARMA models (here we have a 4-dimensional state) and measure similarity by the Martin distance; still, this is not sufficient to separate motion classes (**b**). Discrimination greatly improves by using hybrid autoregressive models with hidden parameters and measuring distances between their probability distributions (**c**). Further evidence in support of the hybrid model is given by the number of dominant components in the parameter posteriors (14) as estimated by Algorithm 1 on the dataset. In (**d**) we show the weights $\alpha_{k,i}$ of the mixture components in the posterior distributions (23) for each sequence k in the dataset. We can see that in the vast majority of sequences the posteriors exhibit multiple modes, whereas in the case of linear dynamics we would have unimodal posteriors. From (**c**) it appears that the limping and walking gaits are not successfully discriminated. This is not surprising: It is hard to limp on a running treadmill, and the two gaits, as we see in Fig. 5, are very similar indeed

step or simply to the fact that the dynamics of the two gaits are very close. In Fig. 6 we also plot the pairwise Martin distance between Gaussian auto-regressive moving-average (ARMA) models learned from the same data using the approach of Bissacco et al. (2001). An ARMA model is a linear dynamical model that is identified using standard tools from the literature of System Identification, as described in Bissacco et al. (2001). While a linear model can be chosen of order high enough to approximate any second-order covariance sequence to an arbitrary degree, here we limit the order of the model to the same order of our hybrid model, showing the our approach out-performs linear models at equal

Table 1 Comparison of gait classification performance in k-nearest neighbor matching using distances between models. We report the fraction of correct k-nearest neighbor matches in the dataset ($k = 3, 5$ and 7) using the same metrics of Fig. 6: Euclidean distance between AR model parameters, Martin distance between ARMA models, and Wasserstein distance between Hybrid AR parameter posteriors. The first number is the rate of correct matches in the entire dataset, in brackets the fraction of correct matches for respectively walking, running and limping sequences. This data show that modeling hybrid dynamics clearly yields better discrimination between gait classes

Model and Measure	$k = 3$	$k = 5$	$k = 7$
Euclidean distance	0.571	0.575	0.592
between AR	(0.632, 0.709, 0.125)	(0.675, 0.671, 0.125)	(0.693, 0.696, 0.150)
Martin distance	0.665	0.648	0.627
between ARMA	(0.842, 0.683, 0.125)	(0.851, 0.633, 0.100)	(0.851, 0.570, 0.100)
Wasserstein distance	0.781	0.794	0.824
between Hybrid AR	(0.860, 0.987, 0.150)	(0.886, 0.987, 0.150)	(0.939, 0.987, 0.175)

complexity. Equivalently, one could show that our model requires less complexity at equal performance levels.

In Table 1 we show the gait classification performance using k-nearest neighbor on each of the three metrics, for some values of k. It is clear that the added descriptive power of our hybrid models combined with the proposed metric on parameters distributions lead to better discrimination between human gait classes.

4 Discussion

We have presented a technique to perform classification in the space of hybrid autoregressive models that we have used to classify human gaits. We have shown that classification based on a hybrid model yields significant improvements over simple linear systems.

In order to achieve our results, we have devised a novel (approximate) filtering and identification technique for hybrid AR models (this is inspired by a wealth of results available in the literature), and introduced a distance between parameter distributions. This distance is not computable efficiently, so we had to resort to an approximation, which nonetheless showed good performance in our experiments.

Our model has limitations. It can only capture *stationary (quasi-periodic) gaits*. Ideally we would like to recognize transient actions, but doing so in a principled manner is well beyond our scope here. We also assume, somewhat optimistically, that temporal statistics are extracted for us from images. This does not mean that we under-appreciate the difficulty in detecting, localizing, and tracking humans in video. On the contrary, the models we propose can be used to *support* these tasks, eventually. Our inference techniques rely on models that can be inferred from images, and we do not assume that forces or higher-order temporal statistics are available, which would be the case if we were analyzing data for graphics or biomechanics.

References

Agarwal, A., & Triggs, B. (2004). Tracking articulated motion using a mixture of autoregressive models. In *Proc. ECCV* (Vol. III, pp. 54–65). Prague 2004.

Bissacco, A. (2005). Modeling and learning contact dynamics in human motion. In *Proc. CVPR* (pp. 421–428).

BenAbdelkader, C., Cutler, R., & Davis, L. (2004). Gait recognition using image self-similarity. *EURASIP Journal on Applied Signal Processing*, *4*, 1–14.

Bickel, P. J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, *9*, 1196–1217.

Bissacco, A., Chiuso, A., Ma, Y., & Soatto, S. (2001). Recognition of human gaits. In *Proc. IEEE int. conf. on comp. vis. and pattern recognit.*, December 2001.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on PAMI*, *23*(3), 257–267.

Bregler, C. (1997). Learning and recognizing human dynamics in video sequences. In *Proc of CVPR* (pp. 568–574).

Bregler, C., & Malik, J. (1998). Tracking people with twists and exponential maps. In *Proc. of CVPR*.

De Coch, K., & De Moor, B. (2000). Subspace angles and distances between ARMA models. In *Proc. of the int. symp. of math. theory of networks and systems*.

Del Vecchio, D., Murray, R. M., & Perona, P. (2003). Decomposition of human motion into dynamics based primitives with application to drawing tasks. *Automatica*, *39*(12), 2085–2098.

Dowson, D. C., & Landau, B. V. (1982). The Frechet distance between multivariate normal distributions. *Journal Multivariate Analysis*, *12*(3), 450–455.

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Proc. of ICCV*.

Garcia, M., Chatterjee, A., & Ruina, A. (1998). The simplest walking model: stability, complexity and scaling. *Journal of Biomechanical Engineering*, *120*, 281–288.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. In *CVIU* (Vol. 73, pp. 82–98).

Ghahramani, Z., & Hinton, G. E. (1998). *Switching state-space models* (Technical Report). Toronto, Canada.

Greenspan, H., Dvir, G., & Rubner, Y. (2004). Context-dependent segmentation and matching in image databases. *Computer Vision and Image Understanding*, *93*, 86–109.

Gustafsson, F. (2000). *Adaptive filtering and change detection*. New York: Wiley.

He, Q., & Debrunner, C. (2000). Individual recognition from periodic activity using hidden Markov models. In *IEEE workshop on human motion*, USA, December 2000.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, *14*(2), 201–211.

Juang, B. H., & Rabiner, L. R. (1985). Mixture autoregressive hidden Markov models for speech signals. *Transactions on Acoustic Speech Signal Processing*, *33*(6), 1404–1413.

Kale, A., Sundaresan, A., Rajagopalan, A. N., Cuntoor, N., Roy-Chowdhury, A., Krueger, V., & Chellappa, R. (2004). Identification of humans using gait. *IEEE Transactions on Image Processing*, *13*(9), 1163–1173.

Krishnaprasad, P. S., & Brockett, R. W. (1979). A scaling theory for linear systems. *IEEE Transactions on Automatic Control*, *AC-25*(2), 197–207.

Lee, C. S., & Elgammal, A. (2004). Gait style and gait content: bilinear models for gait recognition using gait resampling. In *Proc. automatic face and gesture recognition*, Seoul, Korea, 17–19 May 2004.

Lee, L., & Grimson, W. E. L. (2002). Gait analysis for recognition and classification. In *Proc. automatic face and gesture recognition* (pp. 148–155). 20–21 May 2002.

Little, J. J., & Boyd, J. E. (1998). Recognizing people by their gait: the shape of motion. *Videre*, *1*(2), 1–32.

Ljung, L. (1987). *System identification: theory for the user*. New York: Prentice Hall.

Ma, Y., & Vidal, R. (2005). A closed form solution to the identification of hybrid ARX models via identification of algebraic varieties. In *Hybrid systems comput. and control*.

Ma, Y., Soatto, S., Kosecka, J., & Sastry, S. (2003). *An invitation to 3D vision: from images to geometric models*. Berlin: Springer.

Martin, R. (2000). A metric for ARMA processes. *IEEE Transactions on Signal Processing*, *48*(4), 1164–1170.

Mazzaro, C., Sznaier, M., Camps, O., Soatto, S., & Bissacco, A. (2002). A model (in)validation approach to gait recognition. In *Proc. of the 3DPTV*, June 2002.

North, B., Blake, A., Isard, M., & Rittscher, J. (2000). Learning and classification of complex dynamics. *IEEE Transactions on PAMI*, *22*(9), 1016–34.

O'Brien, J. F., Bodenheimer, R. E., Brostow, G. J., & Hodgins, J. K. (2000). Automatic joint parameter estimation from magnetic motion capture data. In *Proc. of graphics interface 2000* (pp. 53–60). Montreal, Canada, May 2000.

Oh, S. M., Rehg, J. M., Balch, T., & Dellaert, F. (2005). Learning and inference in parametric switching linear dynamic systems In *International conference on computer vision (ICCV 05)* (Vol. 2, pp. 1161–1168). Beijing, China, October 2005.

Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on PAMI*, *22*(8), 831–843.

Pavlovic, V., & Rehg, J. (2000). Impact of dynamic model learning on classification of human motion In *Proc. of CVPR*.

Rubner, Y., Tomasi, C., & Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Proc. of ICCV* (pp. 59–66). Bombay, January 1998.

Sarkar, S., Phillips, P. J., Liu, Z., Vega, I. R., Grother, P., & Bowyer, K. W. (2005). The HumanID gait challenge problem: data sets, performance, and analysis. *IEEE Transactions on PAMI*, *27*(2), 162–177.

Shah, M., & Jain, R. (1999). *Motion-based recognition*. Dordrecht: Kluwer.

Tugnait, J. K. (1982). Detection and estimation for abruptly changing systems. *Automatica*, *18*(5), 607–615.

Veres, G. V., Gordon, L., Carter, J. N., & Nixon, M. S. (2004). What image information is important in silhouette-based gait recognition? In *Proc. CVPR 04*, June 2004.

Wilson, A. D., & Bobick, A. F. (1999). Parametric hidden Markov models for gesture recognition. *IEEE Transactions on PAMI*, *21*(9), 884–900.

Zames, G., & El-Sakkary, A. K. (1980). Unstable systems and feedback: the gap metric. In *Proc. of the Allerton conference* (pp. 380–385). October 1980.

Zelnik-Manor, L., & Irani, M. (2006). Statistical analysis of dynamic actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(9), 1530–1535.