Classifying Human Dynamics Without Contact Forces

Alessandro Bissacco Stefano Soatto

UCLA CSD-TR # 050009

Abstract

We develop a classification algorithm for hybrid autoregressive models of human motion for the purpose of video-based analysis and recognition. We assume that some temporal statistics are extracted from the images, and we use them to infer a dynamical system that explicitly models contact forces. We then develop a distance between such models, and compute an approximation that forms the basis for a simple classifier that explicitly factors out exogenous inputs that are not unique to an individual or her gait. We show that such a distance is far more discriminative than the distance between simple linear systems, where most of the energy is devoted to modeling the dynamics of spurious nuisances such as contact forces.

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 [12, 7, 33]. From a modeling perspective, humans are physical objects interacting in 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 [22]), "quasi-static" (e.g. graphs of transitions between poses [31], cumulative video statistics [4]), or "kinematic" representations [5] 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 [17] who showed that even if we strip the image of all of its 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 paper we concentrate on *dynamics as a perceptual cue for human motion recognition*.[3]

If we agree in viewing humans as dynamical systems, then learning their dynamic characteristics is a system identification task [24]. 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 [36], subspace angles [9], Martin's distance [26]. 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 [20].[4] 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.*

[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 imitators display Charlie Chaplin's walk (different masses, inertias and skeletal configuration, same perceptual dynamics). Furthermore, one single snapshot of such moving dot displays rarely yields much information.

[3] 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.

[4] Note that each of these techniques has been applied to the analysis and classification of human motion ([3] for subspace angles and Martin's distance, [27] for the Gap metric) with encouraging but limited results.

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. [6]), 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, this is vision work indeed. In fact, 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.

1.1 Relation to previous work

The literature on human motion recognition is too broad for us to review here. We will provide a synthetic overview of the main approaches, both for the problem of classification of motion gaits [33] and of identification of people by their gait [31].

The proposed approaches can be classified as model-based [5, 3, 21, 19] representing the motion as the parameters of a model fitted to the data, or holistic [16, 23, 35], where some statistics is 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 [31, 19], optical flow [23], joint angles of an articulated body model with image-based tracking [5, 3, 28], or other spatio-temporal motion descriptors [2, 11, 37]. Then some statistics are computed on the reduced data and pattern recognition techniques such as PCA [2], bilinear models [21], Hidden Markov Models [16, 19], K-Nearest Neighbor [23] or Support Vector Machines [22] are applied to the classification problem.

As we motivated in the introduction, we take the approach of modeling the dynamics of human gaits with hybrid linear models. Inference of the state and model parameters for a switching linear model is, in general, NP complete [34]. While several heuristic algorithms exist, 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 system that is particularly attractive since, for each mode, the optimal estimator can be written as a closed-form function of the data [24]. For hybrid-AR models, recent algebraic approaches to filtering and identification [25] have shown promising results, although they are already too involved for our purpose, and therefore we will derive our own identification algorithm in Sect. 2.2. This is our first contribution.

Our second challenge is to define a distance in the space of hybrid-AR models. To the best of our knowledge, this has only been done once before [10] 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 [10] to a stochastic model yields non-sensical distances that either are non-zero when the two models are identical (eq. 7), or that can be infinity for models that are arbitrarily close in the deterministic sense (eq. 8). We define a novel notion of discrepancy that is very intuitive because it ends up coinciding with the Euclidean distance between the optimal estimators.

Our third challenge is to use such a distance for classification. Since we concentrate on the modeling aspect of this problem, we choose the simplest classifier, nearest-neighbor, although one could take our approach and extend it to more sophisticated ones, e.g. kernel machines. The main achievement of this paper is to show that *the distance between hybrid models that factor out contact forces is far 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 super-class 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.

Note that hybrid models for human and robotic locomotion are not new, see for instance [30], so we do not claim novelty in the use of hybrid models. However, the use of hybrid models for classification, whether for human gaits or more in general, has never been attempted before, to the best of our knowledge.

2 Modeling human dynamics for classification

2.1. Autoregresssive Models

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

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

The equation can be rewritten in normal form:

$$y_t = \varphi_t \theta + e_t \tag{2}$$

$$\begin{array}{rcl} \varphi_t & = & \left[\begin{array}{cccc} y_{t-1} \otimes I_p & y_{t-2} \otimes I_p & \cdots & y_{t-n} \otimes I_p \end{array} \right] \\ \theta^T & = & \left[\begin{array}{cccc} \theta_1^T & \theta_2^T & \cdots & \theta_p^T \end{array} \right] \\ \theta_i^T & = & \left[\begin{array}{c} A_1(i,1) \cdots A_1(i,p) \cdots A_n(i,1) \cdots A_n(i,p) \end{array} \right] \end{array}$$

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

Parameter estimation

Assuming Gaussian prior on the parameters: $\theta \sim \mathcal{N}(\theta_0, P_0)$and given a sequence of observations : $y^N = \{y_1, y_2, \cdots, y_N\}$the posterior distribution of the parameters θ is [24]:

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

where:

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

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

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

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:

$$\hat{P} = \left(P_0 + \frac{\sum_{t=1}^{N} y_t^2}{R}\right)^{-1} = \left(P_0 + (N-1)\frac{\Sigma_y}{R}\right)^{-1} \tag{6}$$

where Σ_y is the sample variance of the measurements. The variance $\hat{P}$ is a measure of the uncertainty we have in the estimated parameters. As we could expect, it decreases as the length N of the observation sequence and the signal-to-noise ratio $\frac{\Sigma_y}{R}$ increase. In the limit $N \to \infty$, the variance $\hat{P}$ becomes zero and the estimate $\hat{\theta}$ is the true value of the parameters.

Model discrepancy (AR)

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

$$\begin{array}{rcl} d_e(\theta_1, \theta_2)^2 & = & E[(\theta_1 - \theta_2)^T (\theta_1 - \theta_2)] = \\ & = & (\hat{\theta}_1 - \hat{\theta}_2)^T (\hat{\theta}_1 - \hat{\theta}_2) + \mathrm{Tr}(\hat{P}_1 + \hat{P}_2) \end{array} \tag{7}$$

Unfortunately, this is not a distance; in particular, it is easy to see that $d(\theta_1, \theta_1) \neq 0$. A second approach is to consider the symmetric Kullback-Leibler divergence (K-L) between the two distributions:

$$KL(p_1 || p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} + p_2(x) \log \frac{p_2(x)}{p_1(x)} dx \tag{8}$$

which for our Gaussians becomes:

$$
\begin{aligned}
KL(\theta_1||\theta_2) &= \frac{1}{2}\mathrm{Tr}\left(\hat{P}_2^{-1}\hat{P}_1 + \hat{P}_1^{-1}\hat{P}_2 - 2I\right) + \\
&+ (\hat{\theta}_1 - \hat{\theta}_2)^T \left(\hat{\Sigma}_1^{-1} + \Sigma_2^{-1}\right)(\theta_1 + \theta_2)
\end{aligned}
\tag{9}
$$

The problem with this approach is that as the variances $\hat{\Sigma}_1, \hat{\Sigma}_2$ go to zero (i.e. the quality of the parameter estimates increase), divergence goes to infinity. This happens because 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 indipendently of how far the distributions are. Such a condition is met for example when we have good estimates from sequences generated by models with different underlying parameters.

To overcome these problems, we define the discrepancy $d(\theta_1, \theta_2)$ between θ_1 and θ_2 as the symmetric Kullback-Leibler divergence between the unit variance distributions $\mathcal{N}(\hat{\theta}_1, I_d)$ and $\mathcal{N}(\hat{\theta}_2, I_d)$:

$$
\begin{aligned}
d(\theta_1, \theta_2) &= KL(\mathcal{N}(\hat{\theta}_1, I_d)||\mathcal{N}(\hat{\theta}_2, I_d)) = \\
&= (\hat{\theta}_1 - \hat{\theta}_2)^T(\hat{\theta}_1 - \hat{\theta}_2)
\end{aligned}
\tag{10}
$$

Clearly this is the squared Euclidean distance between the MAP estimations $\hat{\theta}_1$ and $\hat{\theta}_2$. This definition will become useful later when we will extend the discrepancy (10) to Gaussian mixture distributions. Note that we are not simply taking the Euclidean distance between AR parameters, since θ_i are random variables. We will further elaborate on this later.

2.2. Hybrid Autoregressive Models

In order to properly model contact forces in human motion we follow the approach of [1] in using hybrid models where the switches correspond to ground contacts. However, unlike [1], we intend to use such models for classification, and therefore we introduce a different, and to the best of our knowledge novel, switching autoregressive model. This has some similarity with the Autoregressive HMM proposed in [18], 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 $\pi^m = [\pi_1, \cdots, \pi_m]$. To each state q is associated 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:

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

The main novelty of this model is that it induces a distribution on the AR parameters $\theta^m = \{\theta_1, \cdots, \theta_m\}$. In other hybrid AR systems proposed in the literature [10], the parameters θ are modeled as unknown deterministic values. A learning algorithm is derived to compute the maximum likelihood estimate $\theta^{ML} = \arg\max_\theta p(y^N|\theta)$given an observation sequence y^N. Unfortunately, this method does not provide a natural way to compare two model parameters θ_1, θ_2, and a common solution [10] 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 $p(\theta|y^N)$. This allows us to compute distances between models by comparing the posterior distributions on the parameters (see eq. (10)). We can relate the two approaches by considering the case of flat prior $p(\theta) \simeq$ const. Then the posterior $p(\theta|y^N)$ is proportional to the likelihood $p(y^N|\theta)$, and the maximum likelihood estimate is also the maximum a posteriori $\theta^{ML} = \hat{\theta}$. The distance $d^{ML} = ||\theta_1^{ML} - \theta_2^{ML}|| = ||\hat{\theta}_1 - \hat{\theta}_2||$ measures how far the principal modes of the posterior distributions $p(\theta_1|y^N)$ and $p(\theta_2|y^N)$ are. In the case of hybrid models this solution is suboptimal since the posteriors $p(\theta_i|y^N)$ are multimodal mixtures, as we can see in figure 3, while with the distance d^{ML} we are taking into account only one component.

Parameter Estimation

Given an observation sequence y^N we want to estimate the posterior distribution:

$$
\begin{aligned}
p(\theta|y^N, \Lambda) &= \sum_{q^N} p(\theta|q^N, y^N, \Lambda) p(q^N|y^N, \Lambda) = \\
&= \sum_{q^N} \sum_{i=1}^{m} p(\theta_i|q^N, y^N, \Lambda) p(q = i|q^N) p(q^N|y^N, \Lambda)
\end{aligned}
\tag{12}
$$

where $\Lambda = \{\theta_0^m, P_0^m, R^m, M, \pi\}$ are the model parameters, with $\theta_0^m = \{\theta_{0,1}, \cdots, \theta_{0,m}\}$, $P_0^m = \{P_{0,1}, \cdots, P_{0,m}\}$, $R^m = \{R_1, \cdots, R_m\}$. Similarly to (3), we have that $p(\theta_i|q_j^N, y^N, \Lambda) = G(\theta_i; \hat{\theta}_i, \hat{P}_i)$ are Gaussian, $p(q = i|q^N)$ is the relative frequency of state i in the sequence q^N, and $p(q^N|y^N, \Lambda)$ is the posterior of the hidden states given the observations, can be computed in closed form and will be given in the next section. Unfortunately, marginalizing the hidden states $q^N = \{q_1, \cdots, q_N\}$ is computationally intractable because of the exponential complexity. Therefore, we approximate the solution using a bank of K filters, where each filter is tuned on a segmentation hypothesis q_j^N. At each time t we generate a new hypothesis q_t^N by imposing a jump to the most likely sequence and discarding the less likely ones. We therefore approximate the posterior with:

$$p(\theta|y^N, \Lambda) \simeq \frac{1}{C} \sum_{j=1}^{K} \sum_{i=1}^{m} p(\theta_i|q_j^N, y^N, \Lambda) p(q = i|q_j^N) p(q_j^N|y^N, \Lambda) \tag{13}$$

where $C = \sum_{j=1}^{K} p(q_j^N|y^N, \Lambda)$ and q_j^N are the filter hypotheses. Therefore this approximation is a mixture of a constant number Km of Gaussians. In practice we have duplicate hypotheses (due to permutation of the states) and hypotheses with low posterior, so the effective number of components is smaller (figure 3).

Hidden State Filtering

In order to obtain a good approximation of the posterior (13) we need to estimate the K most probable hidden state sequences $q_1^N, \cdots, q_K^N$ given the measurements y^N:

$$\hat{q}_1^N, \cdots, \hat{q}_K^N = \arg \max_{q_1^N, \cdots, q_K^N} \sum_{i=1}^{K} p(q_i^N|y^N, \Lambda) \tag{14}$$

First we derive a recursive expression for the likelihood $p(y^N|q^N, \Lambda)$:

$$p(y^t|q^t, \Lambda) = p(y_t|q^t, y^{t-1}, \Lambda) p(y^{t-1}|q^{t-1}, \Lambda) \tag{15}$$

which yields (see [14]):

$$p(y_t|q^t, y^{t-1}, \Lambda) = G(y_t; \varphi_t^T \hat{\theta}_{q_t,t-1}, \varphi_t^T \hat{P}_{q_t,t-1} \varphi_t + R_{q_t}) \tag{16}$$

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)):

$$\hat{\theta}_{i,t} = \hat{P}_{i,t} \left(P_{0,i}^{-1} \theta_{0,i} + \sum_{j|q_j = i, j \le t} \varphi_j R_i^{-1} y_j \right) \tag{17}$$

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

From (15) we obtain a recursive equation for the posterior up to time t:

$$p(q^t|y^t, \Lambda) = \frac{K_{t-1}}{K_t} p(y_t|q^t, y^{t-1}, \Lambda) p(q_t|q_{t-1}, \Lambda) p(q^{t-1}|y^{t-1}, \Lambda)$$

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 $p(q_1|q_0, \Lambda) = \pi_{q_1}$. Taking the logarithms and substituting (16) yields the update equation:

$$\begin{aligned} \log p(q^t|y^t, \Lambda) = {} & \log p(q^{t-1}|y^{t-1}, \Lambda) + C + \\ & + \log M(q_{t-1}, q_t) - \frac{1}{2} \log \det \Gamma - \\ & - \frac{1}{2} \left(y_t - \varphi_t^T \hat{\theta}_{q_t,t-1} \right)^T \Gamma^{-1} \left(y_t - \varphi_t^T \hat{\theta}_{q_t,t-1} \right) \end{aligned} \tag{19}$$

where $\Gamma = \left(\varphi_t^T \hat{P}_{q_t,t-1} \varphi_t + R_{q_t} \right)$. To find the most probable state sequences (14) we use a bank of K filters, each matched to a hidden state sequence hypothesis q_i^N, and the posterior $p(q^N|y^N, \Lambda)$ is computed recursively with (19). We initialize the filters at $t = 1$ so that there is at least one hypothesis $q_{i,1} = i$ for each possible initial state value $i = \{1, \cdots, m\}$. Then for each time $t = 2, \cdots, N$ we iterate the following rules to maintain the K hypothesis:

- For each hypothesis q_i^t, compute the posterior loglikelihood $\log p(q_i^t|y^t)$ using (19).

- Extend the hypotheses $q_j^t, j = 1, \cdots, K$ to $t+1$ by assuming no switch: $q_j^{t+1} = \{q_j^t, q_{j,t}\}$
- 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 $\{q_{K+i,t+1}\} = \{1, \cdots, m\} \backslash \{q_{o,t}\}$.
- Cut off the $m-1$ least probable sequences, so only K are left.

This algorithm exploits the finite memory property of our hybrid model. Past data do not contain information on what happens after a switch, therefore only the most likely sequence among all with a switch at a given time has to be considered. The number of filters K determines the quality of the estimates. With $K \geq N$, the algorithm is guaranteed to find the optimal state sequences (14) [14]. In order to improve the performances 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.

Discrepancy between Hybrid AR models

We obtain a discrepancy measure between models by extending to hybrid models the distance (10) between posteriors of autoregressive parameters. Let the posterior distributions of the parameters θ_1, θ_2 be

$$p(\theta_1 | y_1^N, \Lambda) = \sum_{i=1}^{n_1} \alpha_{1,i} G(\theta_1; \hat{\theta}_{1,i}, \hat{P}_{1,i})$$

$$p(\theta_2 | y_2^N, \Lambda) = \sum_{i=1}^{n_2} \alpha_{2,i} G(\theta_2; \hat{\theta}_{2,i}, \hat{P}_{2,i})$$

We define the discrepancy $d(\theta_1, \theta_2)$ between the parameters θ_1 and θ_2 to be the symmetric Kullback-Leibler divergence between the mixtures of unit variance Gaussians $\sum_{i=1}^{n_1} \alpha_1 \mathcal{N}(\hat{\theta}_{1,i}, I_d)$ and $\sum_{i=1}^{n_2} \mathcal{N}(\hat{\theta}_{2,i}, I_d)$:

$$d(\theta_1, \theta_2) = KL \left(\sum_{i=1}^{n_1} \alpha_{1,i} \mathcal{N}(\hat{\theta}_{1,i}, I_d) || \sum_{i=1}^{n_2} \alpha_{2,i} \mathcal{N}(\hat{\theta}_{2,i}, I_d) \right) \tag{20}$$

There is no closed form expression for the Kullback-Leibler divergence between mixtures of Gaussians. For computational reasons we use the efficient approximation proposed in [15], which substituted in (20) gives:

$$d(\theta_1, \theta_2) = \sum_{i=1}^{n_1} \alpha_{1,i} \min_{j \in [1, n_2]} \left((\hat{\theta}_{1,i} - \hat{\theta}_{2,j})^T (\hat{\theta}_{1,i} - \hat{\theta}_{2,j}) + \log \frac{\alpha_{1,i}}{\alpha_{2,j}} \right) \tag{21}$$

This approximation matches each Gaussian component of θ_1 with the closest component of θ_2. Since we normalized the variances to the identity, this discrepancy is a weighted sum of the Euclidean distances between the modes of the two mixtures.

In the next section we will illustrate both the computation of the distance and its use for classification of hybrid dynamical models, which we apply to human motion data.

3. Experiments

Our goal in this research is to recognize human motion based on dynamic signatures. We believe that dynamics contain a significant amount of information: Johansson's stripped-down moving-dot displays [17] can allow one to infer whether the person is young, old, happy, sad, even man or woman, which is information likely not coded in the pose or configuration. 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 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, without restriction to human motion. 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.

In the specific case described in this section, the data used in the experiments 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 60Hz; we used 20 retro-reflective markers on the test subjects at the proximity of the body joint locations and 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 subjects, see table 1 for details. Each sequence is sampled at 60 Hz and is about 6 second-long.

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

Figure 1: *List of motion capture data sequences in the gait dataset. For each subject (first column), number of walking, running and limping sequences collected.*

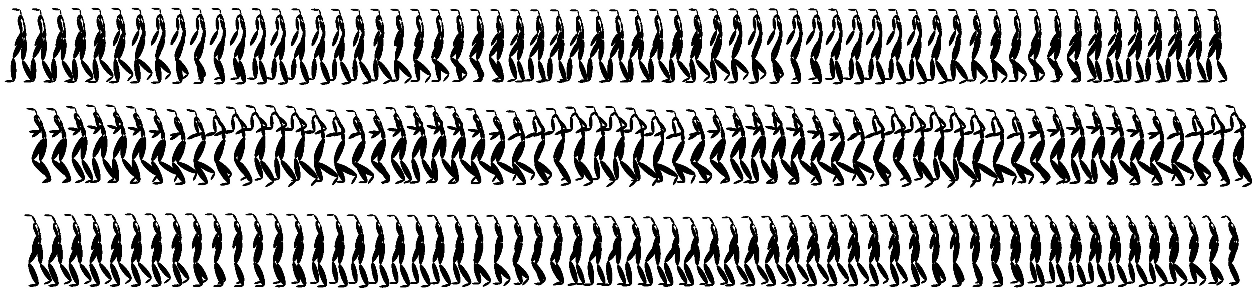

Figure 2: *Short clips (about 3 seconds) from the sequences of the gait dataset. Subject 1 walking (top), running (center) and limping (bottom)*

From marker positions we estimated body skeleton model and join angles with an approach similar to the one proposed in [29]. First we estimate the reference frame moving with the body limb from the set of markers attached to the limb. Then the joint positions are obtained as the center of rotation of the reference frames of adjacent limb. From joint positions, by enforcing fixed limb length we obtained skeleton model and joint angles. Since we do not use a reference model for the skeleton, the estimated skeletons vary from person to person, this affecting the joint angles estimates and making the recognition problem harder. In figure 2 we show some sample clips of the data sequences. Of course we could use pose and configuration information to aid the classification, since that is available in our dataset. However, we are not going to do so because (a) these data are generally not reliable when estimated directly from video, and (b) although pose and configuration are important, many groups are addressing them, and we therefore want to focus our attention on dynamics. Naturally, eventually all will have to be integrated into a complete classification system.
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 expressed in the exponential map parameterization [13]. 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 PCA to each sequence, and retained the first $p = 4$ components. Given a sequence y_i^N we learned the posterior (13) using the algorithm described in section 2.1. The model we propose is very general and contains a number of parameter that should to 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). We set 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, thus modeling the lack of prior information on the parameters. The noise variances R^m are set to the identity, so that in (17) we obtain least squares estimates. We have 2 hidden states and the Markov chain parameters M, π are so that all state have equal probability and the average length of a segment is L: $M(i,i) = \frac{L}{L+1}$, $M(i,j) = \frac{1}{(L+1)(m-1)} \;\; i \neq j$, $\pi_i = \frac{1}{m}$. The posteriors are computed with a bank of K filters. To have optimal segmentations we would need K to be not smaller than the sequence length N, typically about 400. In practice we noticed reducing K to 50 does not change significantly the approximation (13). 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 computing the posterior on the parameters (13). Of all the components of (13), typically only few of them have weight significantly different form zero. Therefore we proceed to pruning all the hypothesis that have weight below a small threshold. In figure (3) we show the histogram of the number of components of the posteriors learned from the gait sequences. We see that most of the sequences have multimodal distribution, with number of modes limited by the number of filters K.

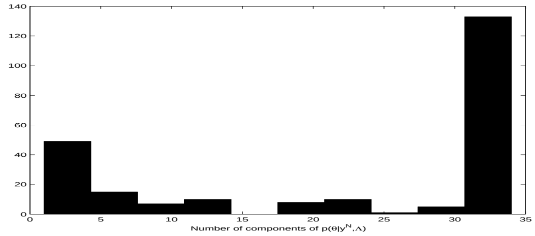

Figure 3: *Histogram of the number of components of the parameter posterior* (13) *for the gait sequences in the dataset. These results show that we cannot assume the posterior to be unimodal.*

3.1 Hybrid models allow for finer dynamic discrimination

The point of this section is to show that hybrid models have more discriminative power than simpler linear models [3], since those must employ most of their modeling assets to capture contact forces that are not specific to an individual or a walking gait. 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, and we do not need to compete with other gait or individual recognition techniques that use different (static) features. Our approach is meant to complement them, not to replace them.

In figure (4) 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 distances (10) between their parameters. By using this simpler similarity measure we would not be able to discriminate between gaits. This is confirmed by the results of a simple nearest neighbor classification: for walking, running and limping we have respectively $37.7\%, 31.7\%, 85\%$ misclassification rates with autoregressive model and $25.4\%, 1.3\%, 85\%$ with hybrid model. The high misclassification rate of limp vs. walk may be due to the different parameterizations of the motion, to the dimensionality reduction step or simply to the fact that the dynamics of the two gaits are very close.

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 in a way that is insensitive to exogenous factors such as contact forces. We have shown that classification based on a hybrid model yields significant improvements.

In order to achieve our results, we had to devise a novel (approximate) filtering and identification technique for hybrid AR models (this is inspired by a wealth or results available in the literature), and introduce a novel distance between such models. This distance is not computable efficiently, so we have proposed an approximation, which we have tested experimentally.

Our results are restricted to *stationary (quasi-periodic) gaits*. Ideally we would like to recognize transient actions, but doing so in a principled manner is beyond our means at the moment, so we prefer to concentrate on a simpler problem. 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, and our inference techniques relies on a model that is inferrable from images. This is not, therefore, a paper in graphics, since it seeks models with discriminative power, not with generative power. 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

[1] Anonymous

[2] C. BenAbdelkader, R. Cutler and L. Davis. Gait Recognition Using Image Self-Similarity. In *EURASIP Journal on Applied Signal Processing*, 2004, Vol 4, pp 1-14.

[3] A. Bissacco, A. Chiuso, Y. Ma and S. Soatto Recognition of Human Gaits. In *Proc. CVPR*, December 2001.

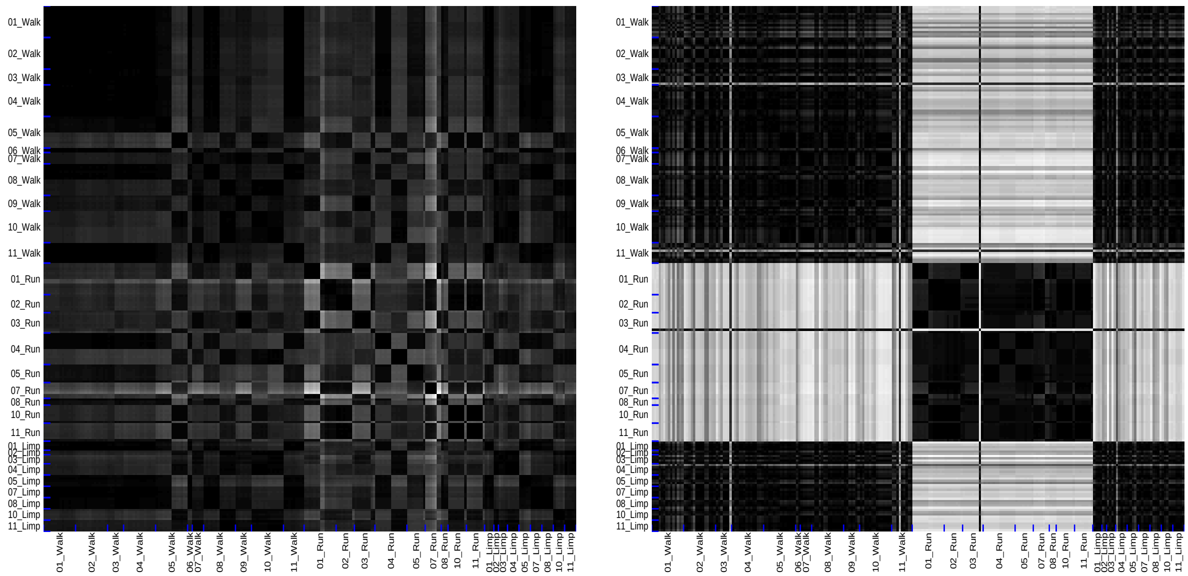

Figure 4: *Discrepancy measure between models learned from the gait dataset. (a) shows the euclidean distance between maximum likelihood estimates of autoregressive model parameters. (b) displays the proposed distance between posterior distributions of the parameters of hybrid autoregressive models. We can see that the simple autoregressive models are not discriminative enough to capture the character of the motion class. Misclassification rates with nearest neighbor for walking, running and limping are respectively* $37.7\%, 31.7\%, 85\%$ *for autoregressive model and* $25.4\%, 1.3\%, 85\%$ *for hybrid model. It appears that the limping and walking gait are not successfully discriminated. This is not surprising: since it is hard to limp on a running treadmill, the dynamics of the two gaits, as we see in figure 2, are very close.*

[4] A. F. Bobick. and J. W. Davis. The recognition of human movement using temporal templates. In *IEEE Trans. PAMI*, 23(3):257-267, 2001.

[5] C. Bregler. Learning and Recognizing Human Dynamics in Video Sequences. In *Proc of CVPR*, pp. 568-574, 1997

[6] C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps In *Proc. of CVPR*, 1998

[7] Q. Cai and J. K. Aggarwal Human motion analysis: a review. In *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, San Juan, Puerto Rico, June 1997.

[8] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. In *IEEE Trans. PAMI*, 22(8), August 2000.

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

[10] D. Del Vecchio, R. M. Murray and P. Perona. Decomposition of Human Motion into Dynamics Based Primitives with Application to Drawing Tasks. In *Automatica*, vol. 39(12), pp. 2085-2098, 2003.

[11] A. A. Efros, A. C. Berg, G. Mori and J. Malik. Recognizing Action at a Distance In *Proc. of ICCV*, 2003.

[12] D. M. Gavrila. The visual analysis of human movement: A survey. In *Computer Vision and Image Understanding*, volume 73, pages 82–98, 1999.

[13] F. S. Grassia. Practical parameterization of rotations using the exponential map In *Journal of Graphics Tools*, Volume 3 , Issue 3, Pages: 29 - 48 , March 1998.

[14] F. Gustafsson. Adaptive filtering and change detection. J. Wiley and Sons, 2000.

[15] J. Goldberger, S. Gordon and H. Greenspan. An efficient imge similarity measure based on approximations of KL-divergence between two gaussian mixtures. In *Proc. ICCV*, 2003

[16] Q. He and C. Debrunner. Individual recognition from periodic activity using Hidden Markov Models. In *IEEE Workshop on Human Motion*, pp. 47-52, Austin, USA, December 2000.

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

[18] B. H. Juang and L. R. Rabiner. Mixture autoregressive hidden Markov models for speech signals. In *IEEE Trans. Acoustic Speech Signal Processing*, vol. ASSP-33, no. 6, pp. 1404-1413, Dec 1985.

[19] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. Cuntoor, A. RoyChowdhury, V. Krueger and R. Chellappa. Identification of Humans Using Gait. In *IEEE Trans. Image Processing*, Vol 13, Issue 9, Sept. 2004, pages 1163-1173. Kluwer, 1997.

[20] P.S. Krishnaprasad and R. W. Brockett A Scaling Theory for Linear Systems. In *IEEE Trans. on Automatic Control*, vol. AC-25, no. 2, pp. 197-207.

[21] C. S. Lee and A. Elgammal. Gait Style and Gait Content: Bilinear Models for Gait Recognition Using Gait Resamplig. In *Proc. Automatic Face and Gesture Recognition*, Seoul, Korea, May17-19, 2004.

[22] L. Lee and W. E. L. Grimson. Gait Analysis for Recognition and Classification. In *Proc. Automatic Face and Gesture Recognition*, May 2002, 20-21, pp. 148-155.

[23] J. J. Little and J. E. Boyd. Recognizing people by their gait: the shape of motion. Videre, 1(2), 1998.

[24] L. Ljung. *System Identification: theory for the user*. Prentice Hall, 1987.

[25] Y. Ma and R. Vidal. A Closed Form Solution to the Identification of Hybrid ARX Models via the Identification of Algebraic Varieties. In *Hybrid Systems Computation and Control*, 2005.

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

[27] C. Mazzaro, M. Sznaier, O. Camps, S. Soatto and A. Bissacco. A Model (In)validation approach to gait recognition. In *In Proc. of the 3DPTV*, June 2002.

[28] B. North, A. Blake, M. Isard and J. Rittscher. Learning and classification of complex dynamics. In *IEEE Trans. PAMI*, volume 22(9), pages 1016-34, 2000.

[29] J. F. O'Brien, R. E. Bodenheimer, G. J. Brostow and J. K. Hodgins. Automatic joint parameter estimation from magnetic motion capture data. In *Proc. of Graphics Interface 2000*, Montreal, Canada, pp. 53-60, May 2000.

[30] F. Plestan, J. W. Grizzle, E. R. Westervelt, and G. Abba. Stable Walking of a 7-DOF Biped Robot. In *IEEE Trans. Robotics and Automation*, Vol. 19, No. 4, pp. 653-668, Aug. 2003.

[31] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother and K. W. Bowyer. The HumanID Gait Challenge Problem: Data Sets, Performance, and Analysis. In *IEEE Trans. PAMI*, Vol. 27, No. 2, February 2005.

[32] H. Sidenbladh and M. J. Black. Learning the statistics of people in images and video. In *International Journal of Computer Vision*, 54(1-3):183-209, 2003.

[33] M. Shah and R. Jain. Motion-Based Recognition. Kluwer, 1999.

[34] J. K. Tugnait. Detection and Estimation for Abruptly Changing Systems. In *Automatica*, Vol. 18, No. 5, pp. 607-615, 1982.

[35] G.V. Veres, L. Gordon, J.N. Carter and M. S. Nixon What image information is important in silhouette-based gait recognition? In *Proc. CVPR 04*, June 2004, Washington D.C., USA.

[36] G. Zames and A. K. El-Sakkary. Unstable systems and feedback: The gap metric. In *Proc. of the Allerton Conference*, pp. 380-385, Oct., 1980.

[37] L. Zelnik-Manor and M. Irani Event-based video analysis. In *Proc of CVPR*, 2001.