MODELING HUMAN GAITS WITH SUBTLETIES

Alessandro Bissacco * **Payam Saisan** **
Stefano Soatto *

* *Department of Computer Science, University of California, Los Angeles, CA 90095*
{bissacco,soatto}@cs.ucla.edu
** *Department of Electrical Engineering, University of California, Los Angeles, CA 90095* saisan@ee.ucla.edu

Abstract: We present a novel approach to modeling subtleties in human motion. We represent the trajectories of a certain number of salient features on the human body as the output of a dynamical system driven by an unknown stochastic input. We present several techniques for inferring model parameters and input signal distributions corresponding to different optimality criteria, and evaluate the corresponding models for accuracy and predictive power. In particular we exploit the higher order statistical information content in motion data to arrive at input signals with independent components and show that the human motion synthesized from non-Gaussian inputs capture best the subtle complexities of the motion data.

Keywords: Image-based modeling and rendering, higher-order statistics, dynamic ICA, human motion modeling, animation

1. INTRODUCTION

Human gaits can convey rich and subtle information. By looking at a person walking from afar, we can sometimes tell whether she is happy, tired, or wounded. We can often make accurate guesses as to individual traits such as the rough age or gender, or even identify a known individual from her gaits. It is clear that such information is encoded not in each individual static pose, but in the *dynamics* of the moving body: in Johansson's experiments Johansson (1973) one cannot tell much from a single frame, but when the sequence is animated suddenly the event becomes easy to parse and identify [1].

Modeling the subtleties in human motion can play a crucial role in a number of applications ranging from security (recognizing individuals from their gaits) to entertainment and the arts (motion capture and synthesis). For the sake of concreteness, we take as our driving application image-based modeling for the purpose of motion synthesis. The idea is simple: we want to collect motion-capture data [2] for an individual, and from these data build a model that can then be used to generate (novel) synthetic motions, for instance of an animated character.

[1] Naturally there is a great deal of information in the photometry and geometry of the scene that can be conveyed in a single static frame. However, in this study we concentrate on the scene dynamics. Johansson's experiments are enlightening because they show that even after stripping a sequence of all of its pictorial content, the dynamics of moving dots still retains information a remarkable amount of information.

[2] In particular, trajectories of a collection of marker positions in space.

In order to do so, we define a gait as the output of a dynamical system driven by a stochastic process with an unknown distribution. While this was done in A. Bissacco and Soatto (2001) for the case of Linear Gaussian models, in this work we generalize the model to arbitrary input distributions. Learning a model then amounts to inferring the model dynamics as well as the input distribution. Unlike the Gaussian case, no closed-form exists, and even the optimality criterion is somewhat open to discussion. In this work we explore several alternatives including Gaussian mixtures, resulting in various mixture Kalman filter models for inference, to exponential classes of densities, resulting in a dynamic version of independent component analysis (ICA) Comon (1994).

The quality of the model inferred can be measured by the size of the residual, for instance the Kullbach-Leibler divergence or the L^2 norm. That measures how well the model fits the training data. However, more importantly for the driving application we are considering, one can simulate the model forward, and visually inspect whether the resulting simulation captures the "character" of the input set. More quantitatively, one can use a portion of the data to learn a model, use the model to predict future data, and then compare that to the real data acquired in subsequent times.

Before delving into the model, we discuss how our approach relates to the state of the art.

1.1 Relation to previous work and contribution of this paper

Modeling subtleties in human or animal motion has been subject of considerable attention lately. Bregler and coworkers Torresani *et al.* (2001) have proposed a variety of methods to model non-rigid motions in an attempt to capture subtleties. Such models are built as linear combinations of a collection of "key" poses, learned using principal component analysis from motion capture data. Similar ideas were used by Brand in his morphable models (see Brand (2001) and references therein). These approaches are also related to a linear-Gaussian model that is a special case of what we describe below. A. Bissacco and Soatto (2001) used a similar model for the purpose of recognition, and defined a metric on the space of model that was used for classification. To the best of our knowledge, dynamical systems with arbitrary input distributions have never been used to model human gaits before.

Local representation of motion based on optical flow has been exploited in Black (1999); Little and Boyd (1996), and view-based methods are proposed in Bobick (1996); Black (1996); Giese and Poggio (2000). Other approaches are based on principal component analysis Yacoob and Black (1999), parameterization of the motion on joint angles Campbell and Bobick (1995) and snake fitting Niyogi (1994). Estimation of motion from stereo Wren (1998) and multiple view systems Gavrila and Davis (1996) have also been investigated. In Bregler (1997) a mixed-state statistical model for the representation of motion has been proposed. In this Switching Linear Dynamic Model a stochastic finite-state automata at the highest level switches between local linear Gaussian models. Estimation and recognition is performed with Expectation-Maximization approaches using particle filters North *et al.* (2000); Black and Jepson (1998) or structured variational inference techniques Pavlovic *et al.* (2000).

Our models are discrete-time, continuous-state dynamical systems, and the action is coded in the dynamical model (i.e. the system parameters) as well as the input distribution. We assume that, using whatever method of the ones described above, either the joint angles or the marker trajectories are given to us. Our work, therefore, comes at a level of abstraction higher than typical motion detection and tracking algorithms. It uses track data to infer a global model of the dynamics of a human gait, using dynamical systems with unknown input distribution.

2. MODELING SUBTLETIES

We assume that we are given the trajectory of a number of distinctive feature points on the human body. For instance, when a motion capture system is used for data acquisition, these points correspond to the position of a number of markers placed ad-hoc on the human subject, for instance at her joints. We denote the position of each marker i ad the time instant t by $y_i(t) \in \mathbb{R}^3$. Alternatively, one may consider the joint angles corresponding to a skeletal model of the subject. In that case, one may call the joint angle i at time t $y_i(t) \in [0,\ C)$, where C_i is a constant that can be either π or 2π depending upon the joint. In any case, the dataset consists of a trajectory in some M-dimensional space; for simplicity we consider the first case where

$$y(t) \in \mathbb{R}^{3M},\ t \in [0,\ t_f] \tag{1}$$

However, the considerations developed here can be transposed to any other representation of the input data one chooses.

As we anticipated, we model $y(t)$ as the output of a dynamical system driven by a stochastic input. That is, we assume that at each instant of time there exists a vector $x(t) \in \mathbb{R}^N$, and suitable functions f and h such that

$$x(t+1) = f(x(t)) + v(t);\ x(0) = x_0 \quad (2)$$

$$y(t) = h(x(t)) + w(t) \quad (3)$$

where x_0 is an unknown but constant vector (the initial condition), and $v(t)$ and $w(t)$ are white, zero-mean stochastic processes. While we assume that $w(t)$ is normally distributed, $w(t) \sim \mathcal{N}(0, R)$, we allow $v(t)$ to be a sample from an unknown probability density q:

$$v(t) \overset{IID}{\sim} q(v). \quad (4)$$

Our goal, then, is to infer the input density q, the initial condition x_0, the order of the model N, the dynamics f and the output map h from a time series $\{y(t)\}_{t\in[0,\ t_f]}$. Naturally, there are many ways of doing so, depending on what inference criterion one chooses. In the following section we describe several criteria that we have evaluated, and the resulting learning algorithms. For the sake of simplicity we restrict our attention to linear dynamics $f(x) = Ax$ and linear output maps $h(x) = Cx$, shifting the emphasis of the modeling power to the input distribution q. Some of the techniques outlined below can be extended to arbitrary nonlinear models, although this is beyond the scope of this paper.

3. INFERENCE CRITERIA AND LEARNING ALGORITHMS

Given a sequence of output data $\{y(t)\}_{t\in[0,\ t_f]}$ generated by a model of the form

$$\begin{aligned} x(t+1) &= Ax(t) + v(t) \\ y(t) &= Cx(t) + w(t) \\ x(0) &= x_0;\ v(t) \overset{IID}{\sim} q(v) \end{aligned} \quad (5)$$

Our goal is that of finding the *optimal* estimates of the unknowns

$$\hat{A}, \hat{C}, \hat{q}(\cdot), \hat{x}(t). \quad (6)$$

We can choose several optimality criteria, depending on whether we seek for the maximum likelihood parameters, the maximum a-posteriori (given a prior distribution on the unknown), the parameters that result in a maximally independent estimated input sequence $\hat{v}(t)$, or the parameters that result in the minimum variance of the output error $y(t) - \hat{y}(t)$. There is no right or wrong criterion; all that one can do is to test several, and compare the results. As we show in the experimental section, comparison can be performed by verifying how well the model captures the training data (residuals), or how well the model can predict future data (prediction error). In addition, for the given application, we can simulate the model and visually inspect the results to verify whether subtleties have indeed been captured.

3.1 Linear Gaussian models and maximum likelihood

If we assume that the input distribution is Gaussian, $q(\cdot) \in \mathcal{N}(0, Q)$, then there are closed-form solution for the identification of the model parameters A, C, Q, R and the state sequence $x(t)$ that minimize the likelihood of the output error. In particular, subspace identification algorithms can be used for this purpose. Since the emphasis of this paper is on non-Gaussian modeling, we refer the reader to Overschee and Moor (1993) for details, and to the experimental section for experiments.

3.2 Dynam-ICA: dynamic independent component analysis

An optimality criterion for inference of model and input descriptions can be constructed by requiring the estimated input sequence $\hat{v}(t)$ to be a realization from a stochastic process that has maximally independent components. Independence can be expressed in terms of the mutual information among input components, which in turn can be written in terms of the Kullback-Leibler divergence.

This approach results in a semi-parametric statistical inference problem, where one has to simultaneously infer the (finite-dimensional) model parameters as well as the (infinite-dimensional) input distribution q. This is essentially an independent component analysis (ICA) problem.

In its conventional static form, ICA attempts to decompose a random vector into a linear combination of statistically independent components. If we call $y \in \mathbb{R}^m$ the random vector, then ICA looks for a matrix $C \in \mathbb{R}^{m\times n}$ with $n \leq m$ and a random vector $x \in \mathbb{R}^n$ with independent components, $p_{\mathbf{x}}(x_1, \ldots, x_n) = p_1(x_1) \ldots p_n(x_n)$ such that

$$y = Cx. \quad (7)$$

The unknowns C and p_i can be estimated by minimizing the mutual information $I(y||Cx) \doteq \int p_{\mathbf{y}} \log \frac{p_{\mathbf{y}}}{p_{Cx}} dy$, computed or approximated using a number of independent and identically distributed (IID) samples from $p_{\mathbf{y}}$: $y(1), \ldots, y(t) \overset{IID}{\sim} p_{\mathbf{y}}$. Typically the process y is assumed to be ergodic, and therefore a time series is used in lieu of a fair sample.

What we have here, however, is a dynamic ICA problem of separating independent components mixed by linear dynamical (state-space) system.

Let us rewrite the output of the model at time t:

$$y(t) = \left[CA^t,\ CA^{t-1}B, \ldots, CB\right] \begin{bmatrix} x(0) \\ v(0) \\ \vdots \\ v(t-1) \end{bmatrix} \doteq \tilde{\mathcal{C}}^t \tilde{\mathbf{V}} \quad (8)$$

and stack the observations $y(1), \ldots, y(t)$ into a vector $\mathbf{Y}^t$ to obtain

$$\mathbf{Y}^t = \tilde{\mathcal{C}}^t \tilde{\mathbf{V}}. \quad (9)$$

One may be tempted to invoke the independence of the components of $\mathbf{V}$ – based on the assumptions that $v(t)$ is white (time samples are independent) and has independent components – and use standard ICA to estimate the mixing matrix $\tilde{\mathcal{C}}^t$. This, however, does not work because it is not possible to use time realizations as independent samples of $\mathbf{Y}$ due to the initial condition $x(0)$.

In what follows, we make the simplifying assumption that the initial condition x_0 is zero. One may conjecture that if t is large enough and A is stable the effect of initial condition will wane; therefore, the assumption may not be as restrictive. Under this assumption, the problem of dynamic ICA can be posed as follows. Consider $\mathbf{Y}^t(k) = [y((k-1)t)^T, \ldots, y(kt-1)^T]^T$, and similarly for $\mathbf{V}^t(k)$. Furthermore, let $\mathcal{C}^t$ be the matrix obtaining by completing, in the sense of Toeplitz, the following matrix

$$\begin{bmatrix} CB & & & \\ CAB & CB & & \\ \vdots & \vdots & \ddots & \\ CA^{t-1}B & CA^{t-2}B & \ldots & CB \end{bmatrix}. \quad (10)$$

Then $\hat{A}, \hat{B}, \hat{C}$ can be found sub-optimally by first estimating the mixing matrix $\mathcal{C}^t$ having the particular structure above from a set of independent samples $\mathbf{Y}^t(1), \ldots, \mathbf{Y}^t(k)$ (notice that $\mathbf{Y}^t(i)$ and $\mathbf{Y}^t(j)$ do not share components $y(k)$):

$$\hat{\mathcal{C}}^t(A, B, C) = \arg\min_{\mathcal{C}^t} I(\mathbf{Y}^t(i) || \mathcal{C}^t \mathbf{V}^t(i)) \quad (11)$$

Under the assumption of stationarity, the model above can be thought of input-output form as $\mathbf{Y}^t(i) = \mathcal{C}^t \mathbf{V}^t(i)$. This is the familiar form of the blind source separation problem. We can reformulate this into an optimization problem where parameters of the model $\mathcal{C}^t_{A,B,C}$ can be learned using Amari's natural gradient flow Amari (1998). A detailed derivation of the parameter learning algorithm for this deconvolution problem is presented in Zhang. and Cichock (1998), where a second stage linear state-space demixing model is used to estimate the input signal with independent components.

In the next section we test this paradigm and learn the model parameters to recover the independent input signals deriving a dynamical system, namely the joint angles for human subjects during walking and running sequences.

We put the algorithm to test by synthesizing a new sequence given the dynamical system parameters of the system and non-Gaussian statistics. Once a model has been inferred, synthesis can be performed by simulating the model forward, that is by sampling an input from the distribution $\hat{q}$, and using it to compute the one-step increment of the state $\hat{x}(t+1)$ and hence the output trajectory $\hat{y}(t)$. In some cases, an explicit expression of the density q may be available. In some other cases, when the order of the model N is very high, sampling may be non trivial, and techniques such as Gibbs sampling may be necessary Geman and Geman (1984). One can apply the state-space demixing system described in Zhang. and Cichock (1998) to the training data to estimate the input distribution $\hat{q}$. The histogram of the values assumed by the elements y_i of $y(t)$ can be used as a discrete approximation $\hat{q}_i$ of the input density. Then synthesis can be easily performed by drawing samples from estimated distributions $\hat{q}_i$.

4. EXPERIMENTS

In this section we learn a model of the form (7) and the statistics of the input v for a number of human motion data sequences corresponding to different subjects' walking and running. The data entails the trajectories of four joint angles corresponding to shoulder, elbow, hips and knees. A typical sequence was about 48 to 78 frames, obtained by applying standard image-based tracking techniques (Bregler (1998)) to video streams taken with a 30Hz camera.

We obtained the dynamical system parameters and estimated the input sequence $v(t)$ with independent components. We then generated a random sample corresponding to the distributions obtained from $v(t)$ to form the input for synthesized sequences. Together with the estimates of the system matrices, A, B and C, we generated new sequences of joint angles. In our preliminary run of the experiments, we implemented a simplified version of the learning algorithm. That is, we first computed the system parameters, matrices A and C using known system identification techniques such as subspace ID and expectation maximization. We then used these to compute the innovations $e(t) = x(t+1) - Ax(t)$. With $e(t) = Bv(t)$ we learned the mixing matrix B using the standard ICA formulation and recovered maximally independent components, $v_i(t)$. We then sampled from the non-gaussian density

function of each $v_i(t)$ to form new input sequences, $\hat{v}_i(t)$. Figure 1 shows the innovations $e(t)$ and the prediction error $e(t) - B\hat{v}(t)$ for one side of the body, the time progression of shoulder, elbow, hip and knee angles. As can be seen in the figures the best results were obtained with the infomax ICA.

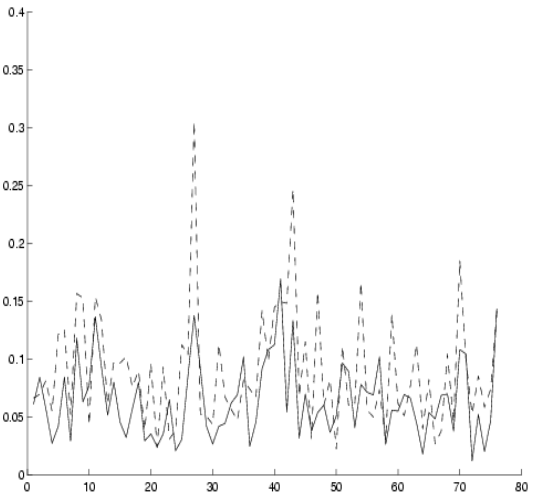

Fig. 1. *Norm of innovations $\|e(t)\|$ (solid line) and prediction errors $\|e(t) - B\hat{v}(t)\|$ (dashed line) normalized with respect the state $\|x(t)\|$. Horizontal axis is time, in frames.*

5. DISCUSSION

We described a novel approach for representing the process of human gaits, in this case walking and running, as the output of a linear dynamical system driven by a stochastic process with independent components. We presented a method for identifying a model from data when transient effects to due initial conditions are neglected. In particular we formulated the model identification process in an information theoretic framework and exploited higher order statistical information content in motion data to form input with independent components. Experimental results suggest the non-Gaussian input models capture best the complexity of the underlying process. We demonstrated that we could use this model to synthesize from it and generate novel instances of different styles of human walking and running motion sequences.

REFERENCES

A. Bissacco, A. Chiuso, Y. Ma and S. Soatto (2001). Recognition of human gaits. *In Proc. of the IEEE Intl. Conf. on Comp. Vision and Patt Recog. pages 401-417.*

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation, 10:251–276.*

Black, M. J. (1996). Eigentracking: robust matching and tracking of articulated objects.

Black, M. J. (1999). Explaining optical flow events with parameterized spatio-temporal models. *In Proc. of Conference on Computer Vision and Pattern Recognition, volume 1, pages 326–332.*

Black, M. J. and A. D. Jepson (1998). A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. *In Proc. of European Conference on Computer Vision, volume 1, pages 909-24.*

Bobick, A. F. (1996). Appearance-based representation of action.

Brand, M. (2001). Morphable 3d models from video. *In Proc. International Conference on Computer Vision and Pattern Recognition.*

Bregler, C. (1997). Learning and recognizing human dynamics in video sequences. *In Proc. of the Conference on Computer Vision and Pattern Recognition, pages 568–574.*

Bregler, C. (1998). Tracking people with twists and exponential maps. *In Proc. International Conference on Computer Vision and Pattern Recognition.*

Campbell, L. and A. Bobick (1995). Recognition of human body motion using phase space constraints. *In Proc. IEEE Conf. on Comp. Vision and Pattern Recogn. page 8.*

Comon, P. (1994). Independent component analysis, a new concept. *Signal Processing, 36:287–314.*

Gavrila, D. M. and L. S. Davis (1996). Tracking of humans in action: a 3-d model-based approach.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *In IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI–6, pp. 721–741.*

Giese, M. A. and T. Poggio (2000). Morphable models for the analysis and synthesis of complex motion patterns. *In International Journal of Computer Vision, volume 38(1), pages 1264–1274.*

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

Little, J. J. and J. E. Boyd (1996). Recognizing people by their gait: the shape of motion.

Niyogi, A. A. (1994). Analyzing and recognizing walking figures in xyt. *In Proc. IEEE Conf. on Comp. Vision and Pattern Recogn., pages 469–474, Seattle, June.*

North, B., A. Blake, M. Isard and J. Rittscher (2000). Learning and classification of complex dynamics. *In IEEE Transaction on Pattern Analysis and Machine Intelligence, volume 22(9), pages 1016-34.*

Overschee, P. Van and B. De Moor (1993). Subspace algorithms for the stochastic identifica-

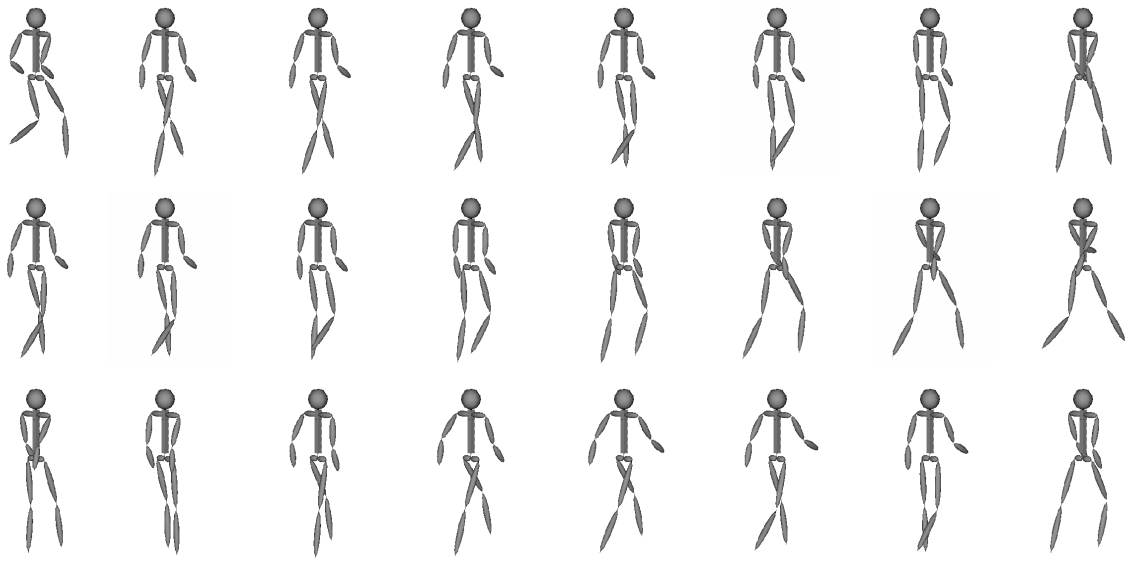

Fig. 2. *Motion sequences. Top row is the original walking data, second row is the synthesized sequence using Gaussian input and third row is the ICA based non-Gaussian input driven sequence.*

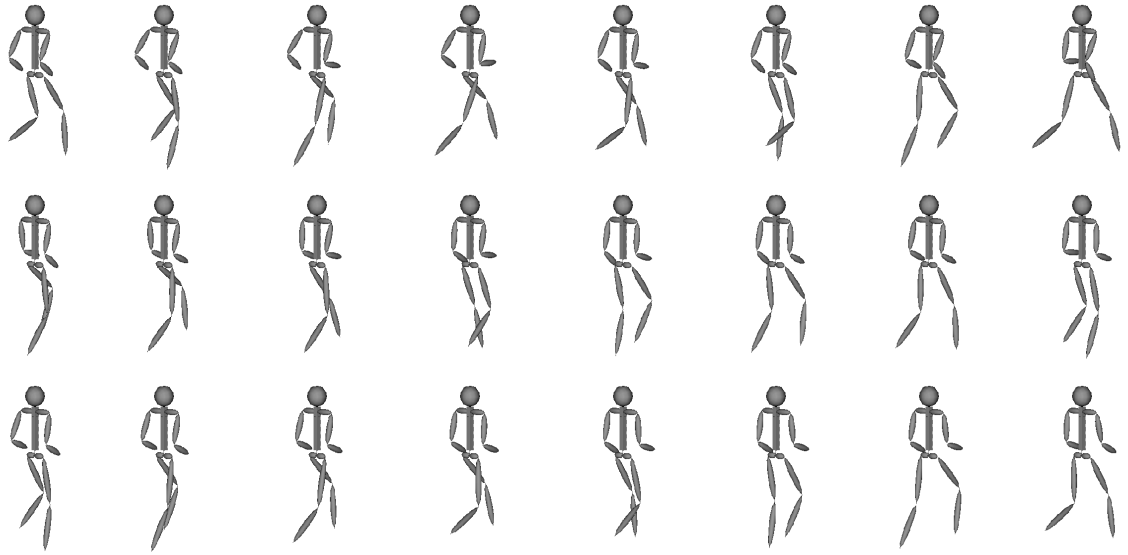

Fig. 3. *Motion sequences. Top row is the original running data, second row is the synthesized sequence using Gaussian input and third row is the ICA based non-Gaussian input driven sequence.*

tion problem. *Automatica, 29:649–660.*

Pavlovic, V., J. Rehg and J. MacCormick (2000). Impact of dynamic model learning on classification of human motio. *In Proc. International Conference on Computer Vision and Pattern Recognition.*

Torresani, L., D. Yang, G. Alexander and C. Bregler (2001). Tracking and modelling non-rigid objects with rank constraint. *In Proc. International Conference on Computer Vision and Pattern Recognition.*

Wren, C. (1998). Dynamic models of human motion.

Yacoob, Y. and M. J. Black (1999). Parameterized modeling and recognition of activities. *In Computer Vision and Image Understanding, volume 73(2), pages 232–247.*

Zhang., L. and A. Cichock (1998). Blind deconvolution of dynamical systems : A state-space approach. *Proceedings of the IEEE. Workshop on NNSP'98, 123–131.*