A Framework for Modeling the Appearance of 3D Articulated Figures

Hedvig Sidenbladh * **Fernando De la Torre** † **Michael J. Black** ‡

*Royal Institute of Technology (KTH), CVAP/NADA, S–100 44 Stockholm, Sweden

hedvig@nada.kth.se, http://www.nada.kth.se/~hedvig/

†La Salle School of Eng., Universitat Ramon Llull, Barcelona, Passeig Bonanova 16, 08028 Spain

ftorre@salleURL.edu, http://www.salleURL.edu/~ftorre/

‡Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304 USA

black@parc.xerox.com, http://www.parc.xerox.com/black/

Abstract

This paper describes a framework for constructing a linear subspace model of image appearance for complex articulated 3D figures such as humans and other animals. A commercial motion capture system provides 3D data that is aligned with images of subjects performing various activities. Portions of a limb's image appearance are seen from multiple views and for multiple subjects. From these partial views, weighted principal component analysis is used to construct a linear subspace representation of the "unwrapped" image appearance of each limb. The linear subspaces provide a generative model of the object appearance that is exploited in a Bayesian particle filtering tracking system. Results of tracking single limbs and walking humans are presented.

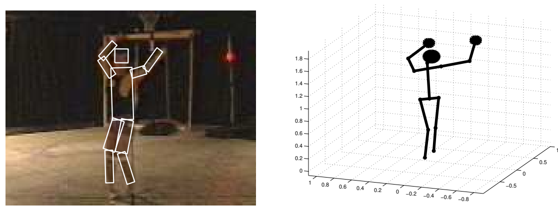

Figure 1. An image from a sequence of a subject dancing, and the corresponding 3D ground truth. For each view, the 3D model is projected into the image, and for each limb, patterns and weights are registered. A linear basis is constructed from this data using weighted principal component analysis.

1 Introduction

The automatic detection and tracking of 3D articulated figures such as humans and other biological creatures in monocular video sequences is a challenging problem with applications in many domains, including human computer interaction, video database search, surveillance, computer graphics, and the scientific analysis of animal behavior. The solution to this problem requires matching a 3D model to image data, but in a monocular image sequence this matching problem is underconstrained. Additionally, models of 3D articulated figures are inherently non-linear and exhibit complex temporal dynamics. To cope with these challenges we work within a Bayesian framework where we represent a probability distribution over the parameters of the 3D model. This framework allows us to exploit recent advances in stochastic search and particle filtering for probabilistic tracking [8, 9] which are robust to singularities and ambiguities in the image appearance. Our approach requires a *generative model* of the object's appearance, the likelihood of observing the image given the model, and a prior distribution over model parameters. In this paper we focus on developing a framework for constructing generative models for the appearance of articulated 3D figures.

Focusing on human figures, we model the body as a collection of articulated cylinders with an associated image representation. Consider, for example, the image shown in Figure 1. A commercial motion capture system is used to gather the "ground truth" 3D motion of the figure. This is used to derive the positions of a 3D cylindrical model of the figure in each image frame. Given the parameters of the camera we can then project each cylinder or limb into the image. The image texture for a particular view of a limb can thus be associated with the cylindrical model. As a person moves we may see their limbs from a variety of views. This can be repeated for multiple subjects, and from this collection of "training views" we build a model for the appearance of a limb.

In general such an appearance model may be quite complex. Humans, for example, wear clothing of highly varied color and pattern. Animals, on the other hand, typically exhibit a limited range of variation in marking and coloration characteristic to the species. Humans too, in certain domains, exhibit limited variability in appearance; for example, sports teams use uniforms with a limited range of patterns. Thus, while in general we may require a non-linear generative model of appearance, we develop the framework here in the context of a linear (eigenspace) model.

Linear subspace methods have been used extensively for constructing appearance models of faces [4, 11, 19] as well as more varied objects [14]. Our problem differs from previous approaches in that we wish to "unwrap" the texture from a roughly cylindrical object and construct a linear basis that represents the full unwrapped appearance. For any given subject, however, we may not see every view of every limb and hence the training set will always be incomplete. Thus, we require a method for constructing a linear subspace representation that takes into account missing data. The visibility of each limb surface is represented by weight maps similar to the cylinder confidence map used by La Cascia and Sclaroff [11]. We exploit the weight maps to perform weighted principal component analysis. The mathematical details of this approach are described and related to recent work in machine learning.

We illustrate how the model can be used to track an arm, as well as a whole walking person in the presence of self occlusion. We define a generative model of image appearance and the likelihood of observing the image given our model. We then briefly outline a temporal prior model for a walking person. A particle filtering method for tracking the person is outlined [3, 9] and results are shown. The focus of this paper is on the framework and mathematics for constructing such models while details of the probabilistic tracking method are described in [17].

2 Related Work

There has been a great deal of work on tracking human heads and bodies in image sequences using models of both shape and appearance (for an overview, see [5]). Methods for full body tracking typically use sparse cues such as background difference images, color (e.g. [20]) or edges [6, 7]. Bregler and Malik [2] tracked a human in 3D using model based motion cues. Motion or optical flow gives rich information, but can cause the tracking model to "drift off" the target. The use of templates [3] avoids this problem, but template tracking is sensitive to changes in view and illumination.

Multiple camera views are often employed to reduce ambiguity and problems due to self occlusion [2, 6]. Although Goncalves *et al.* [7] presented accurate results in tracking an arm in 3D using only one view, there has been little progress in 3D monocular tracking of a whole human body.

Linear subspace methods have been used extensively for modeling, tracking and recognition of faces (e.g. [19]). Edwards *et al.* [4] modeled the shape and greylevel variation of faces independently using principal component analysis (PCA). From this face model they were able to track and identify a face over an image sequence with changes in pose, illumination and expression. Many face tracking and recognition methods model the face as a textured planar surface. In contrast, La Cascia and Sclaroff [11] modeled the head as a cylinder, thus enabling more accurate tracking over wider changes in viewing direction. Given an initial face position, they projected the first image onto the cylinder, creating a cylindrical template used for tracking in subsequent frames. A confidence map was also derived which takes into account the pixel density on the cylindrical template. We use this approach for gathering training data (see Section 3).

Murase and Nayar [14] applied linear subspace methods to objects viewed from multiple orientations. All views of an object were normalized and a single linear subspace was constructed as was the manifold relating the orientation of the view and the coefficients of the model. Black and Jepson [1] used a similar technique to learn appearance models of cylindrical objects but, rather than using a training set with a large number of views, they used a small number of views, and modeled the object as a point in the view eigenspace, plus a linear spatial transformation of the view-based model. Our approach here is quite different. Instead of representing a cylindrical object by a number of views, we wish to "unwrap" the image texture and construct a linear basis in which the basis images represent the full appearance of the object independent of view.

Recently, Cham and Rehg [3] presented a particle filtering approach for human tracking in 2D. Our method is similar, although in 3D. While this increases the dimensionality of the search space, we exploit temporal models of the human motion to constrain the solution to lower dimensional subspaces. In related work, Leventon and Freeman [12] learn a model of short human motion segments from 3D motion capture data. This model is exploited in the probabilistic estimation of 3D human motion given tracking results from a 2D stick-figure model. Here we explore a more constrained temporal model by focusing on human walking and using an approach based on Yacoob and Black's use of "eigencurves" for recognizing human activities [21].

3 Training Data

To construct a generative model from training data we extract example limb patterns from image sequences. Locating the limb positions in the sequences requires correlating the 3D motion capture data with the image data. Below we describe the 3D person and camera models used to compute the transformation from the image space to a limb surface space in which we can represent the limb appearance.

3.1 Coordinate Transformations

The person is modeled as a composite of rigid circular cylinders. Each cylinder is connected to one or several other cylinders with a joint having 1 to 3 degrees of freedom (DOF) represented as Euler angles. The spatial configuration of the person model is determined by the global translation $\mathbf{t}$ and global rotation rotation $\mathbf{r}$ of the cylinder representing the torso of the person model, and the relative Euler angles $\boldsymbol{\alpha}$ between the different cylinders of the model. In total, the model consists of 10 cylinders whose configuration is defined by 25 DOF including the angles at the shoulders, elbows, hips and knees. Hands and feet are not modeled.

A cylinder, or limb, is represented with a radius R, a length L and a homogeneous transformation matrix $\mathbf{T}_l$ which describes the transformation from the global coordinate system to the limb's local coordinate system. $\mathbf{T}_l$ is derived from $\mathbf{t}$, $\mathbf{r}$ and those angles in $\boldsymbol{\alpha}$ relating to limbs connecting limb l and the torso. The local coordinate system is Cartesian with the Z axis directed along the principal axis of the limb (see Figure 2).

We define the limb surface space to be (θ, l) where $\theta \in [0, 2\pi)$ and $l \in [0, L]$. The limb surface space corresponds to cutting the limb open where it crosses the positive X axis and unwrapping it so that $\theta = 0$ (and 2π) at that rotation (see Figure 2).

The camera is modeled as a pinhole camera, with transformation matrix $\mathbf{T}_c$, focal length f and image center $\mathbf{c}$.

We compute the transformation to a point $\mathbf{p}_c$ in the image space from $\mathbf{p}_l = [\theta, l]^T$ in the limb surface space via $\mathbf{P}_l$ in the limb coordinate system and $\mathbf{P}_c = [X_c, Y_c, Z_c, 1]^T$ in the camera coordinate system as:

$$\mathbf{P}_l = [R\cos(\theta), R\sin(\theta), l, 1]^T \quad (1)$$

$$\mathbf{P}_c = \mathbf{T}_c \mathbf{T}_l^{-1} \mathbf{P}_l \quad (2)$$

$$\mathbf{p}_c = \mathbf{c} - f \left[\begin{array}{c} \frac{Z_c}{X_c}, \frac{Y_c}{X_c} \end{array} \right]^T \quad (3)$$

3.2 Acquiring the Limb Pattern Images

Given the model parameters, for each limb we can extract the limb pattern $\mathbf{L}$. Patterns for two different views of a lower right arm is shown in Figure 3.

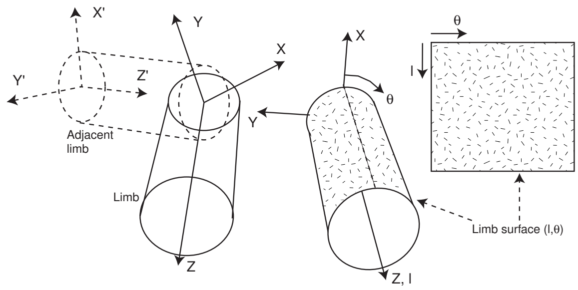

Figure 2. A limb has 1 to 3 rotational DOF. The surface coordinate system (l, θ) of a limb corresponds to cutting the cylinder open where it crosses the positive X axis, and unwrapping it.

Since some parts of the limbs are less visible or not visible at all, we assign a weight to each pixel in each limb surface image indicating its visibility. The weights correspond to the inner product of the surface normal at the limb surface position and the vector from the limb position to the focal point of the camera. This product is positive if the surface faces the camera, and negative otherwise. Since we are not interested in surface points facing away from the camera (they are of course not visible) the weight is set to 0 if the inner product is negative (Figure 3). This weight can also be interpreted as the density of image pixels per limb surface area [11].

Given n example views of a limb, j, let $\mathbf{L}^j = [\mathbf{L}_1^j \ \mathbf{L}_2^j \ ... \ \mathbf{L}_n^j] \in \Re^{d \times n}$, where $\mathbf{L}_i^j$ denotes the image pattern for a particular view, i, of limb j represented as a column vector of length d. For notational simplicity we drop the limb superscript j; models of each limb will be constructed independently. Analogously, for the weight masks, let $\mathbf{W} = [\mathbf{W}_1 \ \mathbf{W}_2 \ ... \ \mathbf{W}_n] \in \Re^{d \times n}$ be the weight images of the limb in column form. For notational purposes, we define $\mathbf{D} = [\mathbf{D}_1 \ \mathbf{D}_2 \ ... \ \mathbf{D}_n] \in \Re^{d \times nd}$ as a matrix where each sub-matrix $\mathbf{D}_i \in \Re^{d \times d}$ is a diagonal matrix with the weights $\mathbf{W}_i$ along the diagonal.

4 Weighted Principal Component Analysis

Principal component analysis (PCA) is a popular statistical tool for performing dimensionality reduction and modeling structure in data [10]. Given the matrix $\mathbf{L}$ we first compute a matrix $\mathbf{A}$ by subtracting the weighted mean

$$\boldsymbol{\mu} = (\sum_{i=1}^{n} \mathbf{D}_i)^{-1} \sum_{i=1}^{n} \mathbf{D}_i \mathbf{L}_i,$$

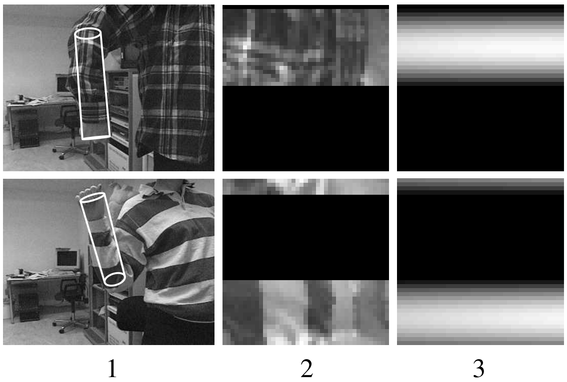

Figure 3. Example of training data. In column 1 we see image with model superimposed, in 2 the extracted limb pattern, in 3 the corresponding weight.

from $\mathbf{L}$. PCA provides the orthogonal transformation $\mathbf{U}$ which minimizes:

$$E(\mathbf{U}) = \sum_{i=1}^{n} ||\mathbf{A}_i - \mathbf{U}\mathbf{U}^T\mathbf{A}_i||^2, \tag{4}$$

where the columns of $\mathbf{U}$ are the k first eigenvectors of the covariance matrix $\mathbf{A}\mathbf{A}^T$. Observe that the matrix $\mathbf{U}$, while not the only basis that spans the subspace of principal components [18, 16], has the numerical advantage of diagonalizing the covariance matrix $\mathbf{A}\mathbf{A}^T$. This means that when the original data in matrix $\mathbf{A}$ has a Gaussian distribution, the projection of the columns of $\mathbf{A}$ on the basis set, $\mathbf{c} = \mathbf{U}^T\mathbf{A}$, will give coefficients that are decorellated and Gaussian. This representation has the advantage of being easy to sample from.

Unlike traditional PCA, we do not have complete data. Given the matrix of textures, $\mathbf{A}$, and the continuous mask, $\mathbf{W}$, the goal is to estimate a linear representation of the image appearance taking into account the observability of the data (i.e. the mask). While singular value decomposition is a common technique for computing a low-dimensional linear model of image data, standard implementations cannot deal with the varying observability as represented by the masks.

Consider instead the probabilistic interpretation of PCA (PPCA) recently proposed independently by Moghaddam and Pentland [13], Tipping and Bishop [18] and Roweis [16]. Tipping and Bishop [18] and Roweis [16] derive an expectation maximization (EM) algorithm for latent variable models, which finds the principal subspace of a set of observed data vectors, assuming an isotropic noise model. That is:

$$\mathbf{A} = \mathbf{B}\mathbf{c} + \boldsymbol{\eta} \tag{5}$$

where $\boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, $\mathbf{c} \sim N(\mathbf{0}, \mathbf{I})$ and $\mathbf{B}$ is a parameter matrix which contains the *factor loadings*.

They showed that the subspace spanned by the principal components can be computed with the EM algorithm when the covariance noise becomes infinitesimal and equal in all the directions; that is, $\lim_{\sigma\to 0} \sigma^2\mathbf{I}$ [16, 18]. In this case the basis vectors can be computed with least-squares optimization, by minimizing the following:

$$\min_{\mathbf{B}} \min_{\mathbf{c}} \sum_{i=1}^{n} ||\mathbf{A}_i - \mathbf{B}\mathbf{c}_i||^2 \tag{6}$$

with respect to the coefficients $\mathbf{c}$ and the basis vectors $\mathbf{B}$. Note that they assume a prior over coefficients $\mathbf{c}$. This would represent a smoothness penalty added to the least square error which acts as a regularization term:

$$\min_{\mathbf{B}} \min_{\mathbf{c}} \sum_{i=1}^{n} (1/\sigma^2)||\mathbf{A}_i - \mathbf{B}\mathbf{c}_i||^2 + \mathbf{c}_i^T\mathbf{c}_i \tag{7}$$

but as $\sigma \to 0$ the smoothness term becomes negligible and the solution becomes the conventional least-squares solution.

Incorporating the weights into (6) the optimization yields:

$$\min_{\mathbf{B}} \min_{\mathbf{c}} \sum_{i=1}^{n} (\mathbf{A}_i - \mathbf{B}\mathbf{c}_i)^T\mathbf{D}_i(\mathbf{A}_i - \mathbf{B}\mathbf{c}_i). \tag{8}$$

As with EM, we alternate between solving for the coefficients, $\mathbf{c}_i$, and the basis vectors $\mathbf{B}$. We have lost, however, the probabilistic interpretation of EM.

The subspace, $\mathbf{B}$, estimated using the above approach is not guaranteed to produce an orthogonal basis set. As with PCA, we would like the basis set $\mathbf{B}$ to be orthogonal as it simplifies sampling from the distribution of coefficients $\mathbf{c}$. In order to compute the orthogonal principal components, we do so iteratively. First, we minimize (8) with a single basis vector in $\mathbf{B}$. Then we subtract from $\mathbf{A}$ the component in the direction of $\mathbf{B}$. To this residual we fit the second basis while imposing orthogonality with respect to the first; e.g. with Gram-Schmidt orthogonalization. This process is repeated until there are enough basis vectors to represent 95% of the variance in the training set. For minimizing (8) we use the Newton-Raphson method. In Figure 4 the mean and basis vectors representing 95% of the variance in a set of 12 views of an arm in 3 different shirts are shown.

5 Tracking

The learned limb bases $\mathbf{B}$ are exploited to perform 3D articulated tracking of people in monocular image sequences. The Bayesian tracking framework is described only briefly

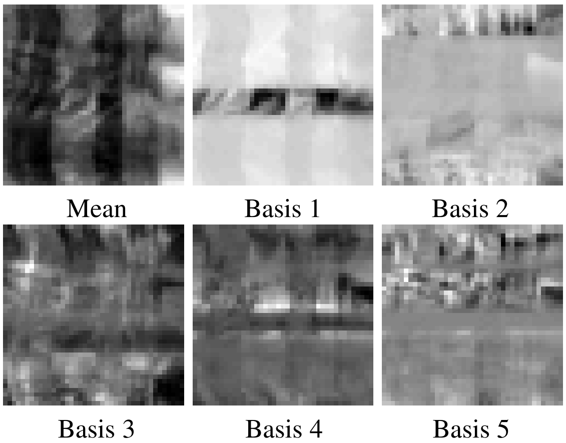

Figure 4. Mean and the first five eigenvectors in a training set of arm views. All images are normalized to greylevel $[0, 255]$ for visibility. The horizontal stripes in the eigenimages occur because each training image is weighted by its corresponding weight image. Thus, different training images have effects on different parts of the eigenimages.

here; a more detailed treatment is presented in [17]. The pose of the human body is defined by the the rotation $\mathbf{r}$ and translation $\mathbf{t}$ of the torso and the relative angles, $\boldsymbol{\alpha}$, of the joints (see Section 3.1). Additionally, the linear coefficients $\mathbf{c}^j$, together with the trained linear subspaces of each limb j, define the image appearance of the limbs. All these parameters together define a *generative model* for the human's appearance in the image.

The generative model can be seen as a "template generator". For a given vector of limb appearance coefficients $\mathbf{c}$, synthesized patterns similar to those in Figure 3 can be generated as $\mathbf{L}_{template} = \boldsymbol{\mu} + \mathbf{Bc}$. For a given pose of the articulated model, we compute the current weight images from the orientation of the limb surfaces with respect to the camera viewing angle (see Section 3.2). We also extract the actual patterns $\mathbf{L}$ on the limbs in the current image. Now in analogy to template matching the difference between $\mathbf{L}_{template}$ and $\mathbf{L}$, weighted by the computed weight image, can be computed.

Let the generative model at a specific time instance t be defined by the set of parameters $\boldsymbol{\phi}_t = [\mathbf{r}_t, \mathbf{t}_t, \boldsymbol{\alpha}_t, \mathbf{c}_t]^T$. Taking a Bayesian approach, the posterior probability of the distribution over $\boldsymbol{\phi}_t$ can be written as [9, 17]:

$$p(\boldsymbol{\phi}_t|\mathbf{I}_t) = K_t\, p(\mathbf{I}_t|\boldsymbol{\phi}_t)\, p(\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}) \qquad (9)$$

where K_t is a normalization constant independent of $\boldsymbol{\phi}_t$ and $\mathbf{I}_t$ is the image at time t.

In the subsequent sections, the different parts of (9) will be discussed. First we describe how the likelihood $p(\mathbf{I}_t|\boldsymbol{\phi}_t)$ is defined, then how the probability distribution over $\boldsymbol{\phi}$ is propagated over time.

5.1 Likelihood

The likelihood of a configuration of the generative model is a measure of how well the image data fits the model. In other words, we want to compare the limb patterns according to the coefficients $\mathbf{c}$ in the generative model with the actual patterns in the image.

Given an image $\mathbf{I}_t$ the parameters in $\boldsymbol{\phi}_t$ are used to extract the mean-subtracted limb patterns $\mathbf{A} = \mathbf{L} - \boldsymbol{\mu}$ and weight matrices $\mathbf{D}$ in the same way as described in Section 3. We define the likelihood in terms of the distance between the observed image patterns $\mathbf{A}$ and the generated patterns given by the limb appearance coefficients $\mathbf{c}$, weighted by $\mathbf{D}$:

$$p(\mathbf{I}_t|\boldsymbol{\phi}_t) = \prod_{j=1}^{m} p^j \qquad (10)$$

$$p^j = \begin{cases} p_{occluded} & \text{if } O^j \\ \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(\mathbf{A}^j - \mathbf{B}^j\mathbf{c}^j)^T \mathbf{D}^j (\mathbf{A}^j - \mathbf{B}^j\mathbf{c}^j)}{2\sigma^2 \mathrm{tr}\mathbf{D}^j}} & \text{if } \neg O^j \end{cases}$$

where m is the number of limbs, O^j is true if limb j is occluded by other limbs, $p_{occluded}$ is the probability of occlusion, σ is the standard deviation of the Mahalanobis distance between the estimated and actual limb patterns, assumed equal for all limbs, and $\mathbf{B}^j$ is the learned basis for limb j.

5.2 Temporal Model

The temporal model defines the probability of observing the body in a certain pose with a particular appearance given its pose and appearance at the previous time instant. This temporal prior can help constrain the distribution over model parameters to regions of the parameter space that are likely to contain the solution. In this paper, we examine two temporal models: a linear model of smooth motion and a model specific to walking [17]. In the smooth motion model, the parameters $\mathbf{r}$, $\mathbf{t}$, $\boldsymbol{\alpha}$ and $\mathbf{c}$ are propagated in time independent of each other:

$$p(\mathbf{t}_t|\mathbf{t}_{t-1}) = G(\mathbf{t}_t - \mathbf{t}_{t-1}, \boldsymbol{\sigma}_t) \qquad (11)$$

$$p(\mathbf{r}_t|\mathbf{r}_{t-1}) = G(\mathbf{r}_t - \mathbf{r}_{t-1}, \boldsymbol{\sigma}_r) \qquad (12)$$

$$p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}) = G(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1}, \boldsymbol{\sigma}_\alpha) \qquad (13)$$

$$p(\mathbf{c}_t|\mathbf{c}_{t-1}) = G(\mathbf{c}_t - \mathbf{c}_{t-1}, \boldsymbol{\sigma}_c) \qquad (14)$$

where $G(x, \sigma)$ is a zero-mean Gaussian with standard deviation σ evaluated at x. $\boldsymbol{\sigma}_t$, $\boldsymbol{\sigma}_r$ and $\boldsymbol{\sigma}_\alpha$ are empirically determined and $\boldsymbol{\sigma}_c = \varepsilon\boldsymbol{\lambda}$ where ε is a small number and $\boldsymbol{\lambda}$

are the eigenvalues corresponding to the basis $\mathbf{B}$. All parameters are initiated manually except $\mathbf{c}$ which is initiated as $p(\mathbf{c}_0) = G(\mathbf{c}_0, \boldsymbol{\lambda})$.

However, in the walking model, dependencies between angles are learned from examples. A commercial motion capture system is used to gather a number of example walking cycles from different individuals. After normalization of the walking cycles with respect to time, each cycle i is represented by a vector $\mathbf{V}_i$ consisting of the cycles of all joint angles $\boldsymbol{\alpha}$ concatenated.

The mean $\boldsymbol{\mu}_V$ of the vectors $\mathbf{V}_i$ is computed and subtracted from the vectors. Then, multivariate principal component analysis [15, 21] is used to learn a basis for walking cycles. The 5 largest eigenmodes $\mathbf{B}_V$ representing 95% of the variance in the training set are selected. The evolution in time of the relative joint angles $\boldsymbol{\alpha}$ is then determined by the eigencoefficients $\mathbf{c}_\alpha$ and some parameter ρ_α determining phase in the walking cycle. In each time instant, $\boldsymbol{\alpha}$ can be computed as $\boldsymbol{\alpha} = (\boldsymbol{\mu}_V + \mathbf{B}_V \mathbf{c}_\alpha)[\rho_\alpha]$. Thus, over time, we only need to propagate $\mathbf{c}_\alpha$ and ρ_α, which have fewer dimensions than $\boldsymbol{\alpha}$ and vary in more predictable ways.

The parameters determining $\boldsymbol{\alpha}$ are propagated in time as:

$$\begin{aligned} p(\mathbf{c}_{\alpha,t}|\mathbf{c}_{\alpha,t-1}) &= G(\mathbf{c}_{\alpha,t} - \mathbf{c}_{\alpha,t-1}, \boldsymbol{\sigma}_{c_\alpha}) & (15) \\ p(\rho_{\alpha,t}|\rho_{\alpha,t-1}) &= G(\rho_{\alpha,t} - \rho_{\alpha,t-1}, \sigma_\rho) & (16) \end{aligned}$$

where σ_ρ is empirically determined and, in analogy with $\boldsymbol{\sigma}_c$, $\boldsymbol{\sigma}_{c_\alpha} = \varepsilon_V \boldsymbol{\lambda}_V$. The phase parameter ρ_α is initiated manually and $\mathbf{c}_\alpha$ is initiated as $p(\mathbf{c}_{\alpha,0}) = G(\mathbf{c}_{\alpha,0}, \boldsymbol{\lambda}_V)$.

5.3 Propagation in Time

The mapping from the parameters $\boldsymbol{\phi}$ of the generative model to the image positions of the limbs is nonlinear and the potentially complex image structure of human clothing results in matching ambiguities. Thus, the likelihood distribution cannot be computed in closed form over the parameter space $\boldsymbol{\phi}_t$. However, it is easy to evaluate the likelihood for a particular value of $\boldsymbol{\phi}_t$. Therefore, instead of computing the posterior distribution analytically at each time step, we chose to use a sampling technique to represent the posterior distribution and propagate it in time [9, 17].

We represent the distribution over $\boldsymbol{\phi}_t$ as a N samples, where N is a large number (see Figure 5). For propagating the distribution in time, we use the Condensation [9] algorithm which is a particle filtering technique.

At each time step, a new set of samples are drawn from the posterior distribution in the previous time step. The new distribution is propagated in time according to one of the temporal models described above. Then, the likelihood of each sample is evaluated. This gives an approximation of the current posterior distribution.

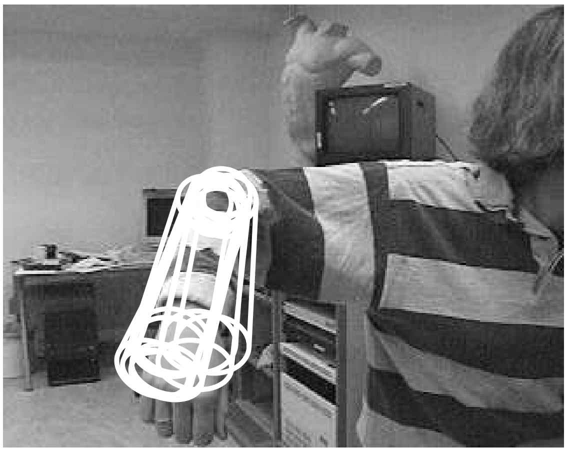

Figure 5. Illustration of a sampled distribution: 10 samples from a posterior distribution over ϕ_t for an arm, projected into the image coordinate system.

6 Results

In order to test the concept of modeling surface structure on cylindrical objects, we implemented the particle filtering algorithm for one arm with a general smooth motion prior, and for the whole body with a walking prior described in Section 5.

6.1 Tracking of Arm

To test the performance of the likelihood measure, we used a model of one cylinder to track a lower arm, using a smooth motion prior. An arm eigenspace was learned from 12 different views of the arm, with 3 different shirts. The result of the tracking can be seen in Figure 6. Distortions in the actual arm pattern compared to the trained pattern are mostly due to wrinkles on the shirt (making it non-cylindrical).

6.2 Tracking of Walking Subject

In this experiment we only modeled the appearance of one subject. Therefore, the likelihood measure (10) was simplified to:

$$\begin{aligned} p(\mathbf{I}_t|\boldsymbol{\phi}_t) &= \prod_{l=1}^{m} p^j & (17) \\ p^j &= \begin{cases} p_{occluded} & \text{if } O^j \\ \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{\mathbf{A}^{j\mathrm{T}} \mathbf{D}^j \mathbf{A}^j}{2\sigma^2 \mathrm{tr}\mathbf{D}^j}} & \text{if } \neg O^j \end{cases} \end{aligned}$$

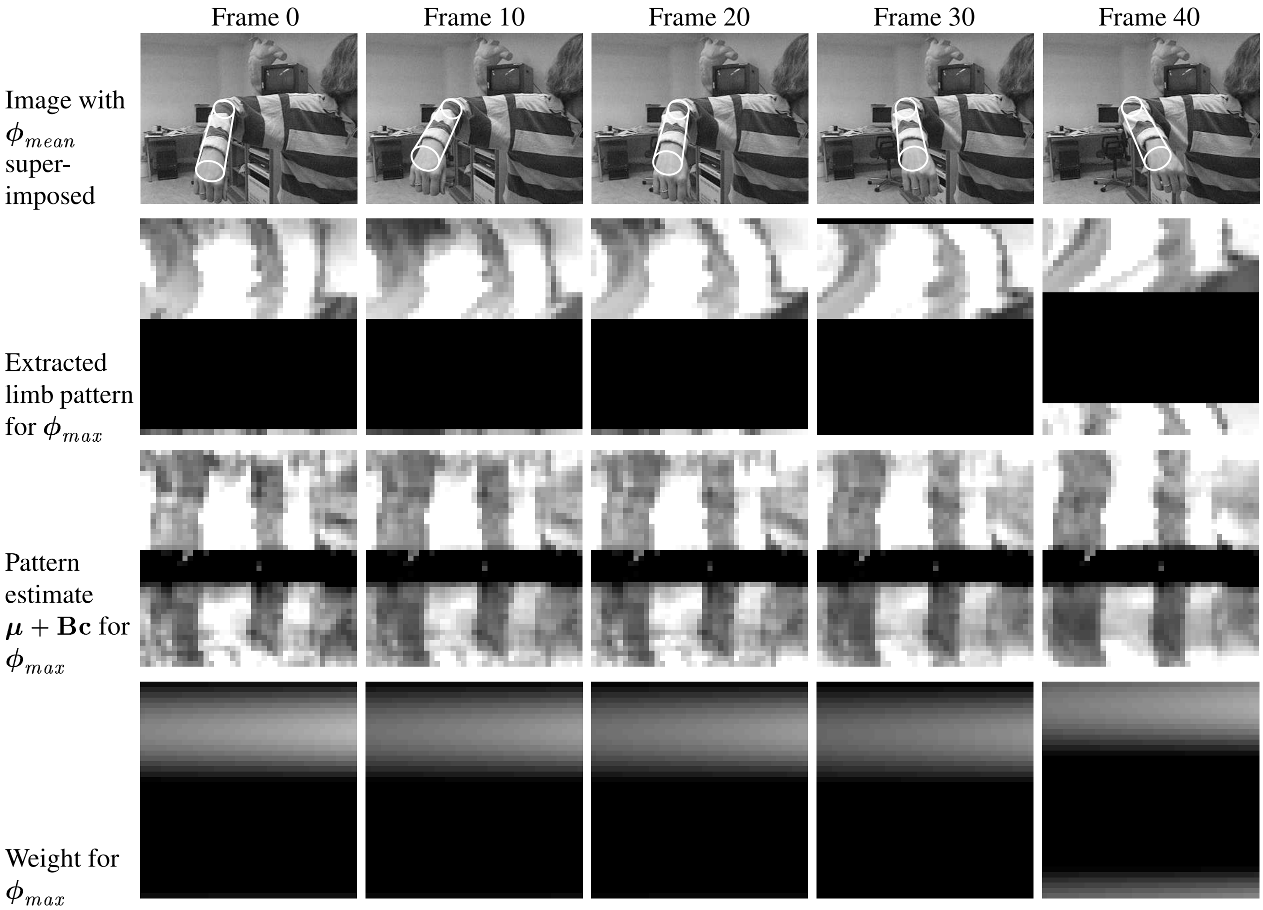

Figure 6. Tracking of a lower arm moving back and forth. In row 1 the image $\mathbf{I}$ is shown, with the weighted mean of the posterior distribution $\phi_{mean} = \frac{1}{N}\sum_{i=1}^{N} \phi_i p(\mathbf{I}|\phi_i)$ superimposed. Row 2 shows the extracted limb pattern $\mathbf{L}$ for the sample ϕ_{max} with the largest likelihood. Row 3 shows the pattern $\boldsymbol{\mu} + \mathbf{Bc}$ for ϕ_{max}, while the weight image for ϕ_{max} is shown in row 4.

This means that we do not take the parameters $\mathbf{c}^j$ into regard. Instead, the likelihood p^j depends on the distance to the learned mean limb image $\boldsymbol{\mu}^j$. This is equivalent to the use of a cylindrical template.

The limb appearances were learned from a number of views of a subject dancing, and then used for tracking the same subject walking in another sequence, using the walking temporal model. The height of the figure in the images is approximately 50 pixels, making the resolution very low. In Figure 7 we see parts of the sequence with the weighted mean $\boldsymbol{\phi}_{mean} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}_i p(\mathbf{I}|\boldsymbol{\phi}_i)$ superimposed.

7 Summary and Discussion

We present a framework for modeling of the appearance of 3D articulated figures composed of cylindrical elements. Given 3D motion capture data of a person moving, correlated with image data, we can construct a linear basis for the appearance of the person's limbs using weighted linear subspace analysis. This approach has the advantage that non-random missing data is explicitly taken into account in the learning of the basis. Thus, we are able to learn a generative model of the appearance of all surfaces on the figure regardless of viewing direction, in contrast to the view-dependent approaches such as template matching.

Modeling an articulated figure as a composite of cylinders is of course limiting since most biological creatures have a reformable structure, and show variability in size and shape. This variability over time and population needs to be modeled as well as the variability in appearance. Instead of cylinders we could use eigenshapes, superquadrics or other high dimensional structures. However, there is a tradeoff between spatial accuracy of the generative model and efficiency in the filtering algorithm. When possible, a simpler model is preferred.

Figure 7. Tracking of a walking subject. The 7 images are frame 0, 5, 10, 15, 20, 25 and 30 of the sequence. Overlaid on the images are the weighted means of the posterior probability distribution.

One drawback of using a linear subspace to model the appearance of humans is that the patterns of modern clothing vary in complex ways. Animals, however, often show characteristic patterns specific to the species. Given 3D motion capture data for several individuals of a species, a generative appearance model could be learned for that species. Reliable tracking of animals has several interesting applications, such as behavior analysis. To represent the complexity of human clothing will require more sophisticated, non-linear, generative texture models.

This work represents a preliminary step towards building generative models of human appearance. Such models may have applications in tracking as shown here as well as in computer graphics. A great deal remains to be done however to construct representations and learning methods for more complex image textures. In our current work we are extending the Bayesian tracking framework with more realistic likelihood models, better temporal models of human motion, the incorporation of additional image cues, and refinements to the tracking algorithm.

Acknowledgements

We are very grateful to Michael Gleicher for generously providing the 3D motion capture data and image sequences. David Fleet, Dirk Ormoneit, and Ruth Rosenholtz helped us work through many of the ideas described here. Special thanks go to Manolis Kamvysselis for mocap data wrangling in Matlab.

References

[1] M. J. Black and A. D. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation. *IJCV*, 26(1):63–84, 1998.

[2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, 1998.

[3] T.-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *CVPR*, volume 1, pages 239–245, 1999.

[4] G. J. Edwards, T. F. Cootes, and C. J. Taylor. Face recognition using active appearance models. In *ECCV*, pages 581–595, 1998.

[5] D. M. Gavrila. The visual analysis of human movement: a survey. *CVIU*, 73(1):82–98, 1999.

[6] D. M. Gavrila and L. S. Davis. 3-D model-based tracking of humans in action: a multi-view approach. In *CVPR*, pages 73–80, 1996.

[7] L. Goncalves, E. Di Bernardi, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3D. In *ICCV*, 1995.

[8] N. Gordon. A novel approach to nonlinear/non-gaussian Bayesian state estimation. *IEE Proceedings on Radar, Sonar and Navigation*, 140(2):107–113, 1996.

[9] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV*, pages 343–356, 1996.

[10] I. T. Jolliffe. *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[11] M. La Cascia and S. Sclaroff. Fast, reliable head tracking under varying conditions. In *CVPR*, pages 604–609, 1999.

[12] M. E. Leventon and W. T. Freeman. Bayesian estimation of 3-d human motion from an image sequence. Technical Report TR–98–06, Mitsubishi Electric Research Lab, 1998.

[13] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *ICCV*, 1995.

[14] H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearance. *IJCV*, 14:5–24, 1995.

[15] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. New York: Springer Verlag, 1997.

[16] S. Roweis. EM algorithms for PCA and SPCA. In *Neural Information Processing Systems*, 1997.

[17] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *In review*.

[18] M. Tipping and C. M. Bishop. Probabilistic principal component analysis. Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, 1997.

[19] M. Turk and A. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–591, 1991.

[20] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: real-time tracking of the human body. *PAMI*, 19(7):780–785, 1997.

[21] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *CVIU*, 73(2):232–247, 1999.