MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Bayesian Estimation of 3-D Human Motion

Michael E. Leventon, William T. Freeman

TR98-06 December 1998

Abstract

We address the problem of reconstructing the 3-dimensional motions of a human figure from a monocular image sequence. We take a statistical approach, and use a set of motion capture examples to build a gaussian probability model for short human motion sequences. We first study this model in a simplified rendering domain. This yields analytic results for the optimal 3-d estimate given a 2-d temporal sequence, as well as for which motion modes are difficult to estimate. The results from the simplified rendering conditions show that if we can overlay a stick figure on an image of a moving human, we can estimate his or her 3-d motion well. We built an interactive tracking system to process real video sequences, and can achieve good 3-d reconstructions of the human figure motion.

MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Bayesian estimation of 3-d human motion from an image sequence

Michael E. Leventon
Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, MA 02139
leventon@ai.mit.edu
William T. Freeman
MERL, Mitsubishi Electric Res. Lab.
201 Broadway
Cambridge, MA 02139
freeman@merl.com
TR-98-06 July 1998

Abstract

We address the problem of reconstructing the 3-dimensional motions of a human figure from a monocular image sequence. We take a statistical approach, and use a set of motion capture examples to build a gaussian probability model for short human motion sequences. We first study this model in a simplified rendering domain. This yields analytic results for the optimal 3-d estimate given a 2-d temporal sequence, as well as for which motion modes are difficult to estimate. The results from the simplified rendering conditions show that if we can overlay a stick figure on an image of a moving human, we can estimate his or her 3-d motion well. We built an interactive tracking system to process real video sequences, and can achieve good 3-d reconstructions of the human figure motion.

1. First printing, TR98-05, April, 1998

1 Introduction

As one watches a film or video of a person moving, one can easily estimate the 3-dimensional motions of the moving person from watching the 2-d projected images over time. A dancer could repeat the motions depicted in the film. Yet such 3-d motion is hard for a computer to estimate.

Many applications would follow from a computer with the same abilities to infer 3-d motions. There are applications to public safety for elevators and escalators, as well as in interactive games, and virtual reality. In computer graphics, a growing industry is devoted to "motion capture", where digitized human figure motion drives computer graphic characters. The human's 3-d motion information is digitized either by magnetic sensors or by optical techniques with multiple calibrated cameras and a special suit of markers. Unfortunately, either technique is expensive and cumbersome. To obtain 3-d figure motion information from single-camera video would allow motion capture driven by ordinary monocular video cameras, and could be applied to archival film or video.

Under constrained viewing and motion conditions, Goncalves and collaborators [8] tracked the motion of an arm in 3-d. Rehg and Kanade [13] track some hand motions over 3-d, allowing significant occlusions. However, this requires 3-d model initialization, and controlled viewing conditions. Work at recovering body pose from more than one camera has met with more success (e.g. [6, 10]). Despite research attention (e.g., [5]), the problem of recovering 3-d figure motion from single camera video has not been solved satisfactorily.

Our approach is to use strong prior knowledge about how humans move. We show that this prior knowledge dramatically improves the 3d reconstructions. We learn our prior model from examples of 3-d human motion.

We first study the 3-d reconstruction in a simplified image rendering domain where a Bayesian analysis provides analytic solutions to fundamental questions about estimating figural motion from image data. Using insights from the simplified domain, we apply our Bayesian method to real images and reconstruct human figure motions from archival video. Our system accomodates interactive correction of automated 2-d tracking errors, which allows reconstruction even from difficult film sequences.

2 Prior model

Our training examples are 10 3-d motion capture sequences, of 5 - 10 seconds each, obtained from `http://www.biovision.com`. The data is position information of 37 markers over 120 to 240 temporal frames for each sequence, sampled at roughly 20 frames per second. The motions are an ecclectic set of short activities, presumably designed to illustrate the range and precision of the motion capture equipment. Fig. 1 shows subsets of 3 of the 10 motion sequences. (As in our other motion displays, these figures show a "stroboscopic" display of an animated sequence, temporally subsampled and sometimes spatially offset for clarity. We draw lines between markers (circles) to create a stick person that is easier to interpret.)

We seek a simple and tractable, yet useful, probabilistic model for the 3d motions, learned from these examples. We divide up the motion signals into "segments" of a fixed, short temporal length. Our prior model will be a probability distribution over those temporal segments, marker positions over a few frames.

If we choose too many frames for our units of human motion, our training data won't be long enough to give us a reliable enough model of such a complex vector. If we choose too few frames, we won't capture enough motion regularities with our model. For what follows, we found 10 to be a good number of frames for each segment. Sampling from the original data, with an overlapped offset of 5 frames, we obtained 257 10-frame segments, represented by 1110 numbers each (37 body markers times 3 dimensions times 10 frames).

We want a probabilistic model that describes these vectors. Motivated by the success of principle components analysis (PCA) at dimensionality reduction [4, 14], we first ask whether we can describe these motion segments as linear combinations of basis functions. We form a training matrix, M, by stacking the 1110 dimensional training vectors together in columns, after first subtracting the mean vector, $\vec{m}$. Singular value decomposition, SVD, gives $M = USV'$, where the columns of U are the basis functions and the diagonal elements of S are the corresponding singluar values. The solid line of Fig. 3 shows the singular value spectrum. The spectrum drops quickly, allowing a good summary of the data (91% of the variance) from just 50 eigenvectors. Figure 2 shows a typical motion sequence synthesized using 40 eigenvectors, showing an imperfect, but good, reconstruction. Thus, we can summarize the 1110 dimensional motion segment vectors by their coordinates in a 50

dimensional subspace of the 1110 dimensional space.

Of course, the singular values themselves provide additional information about the motion segment data. We can model the data as resulting from a gaussian probability distribution of covariance $\Lambda = US$ [4]. This probabilistic model is much stronger than just the subspace information itself (e.g., [12]). Figure 4 shows three random draws from the resulting probability model for 10-frame human motion segments. The motions look like plausible human motions, although, of course, random. This gaussian distribution provides a useful prior model for how a human moves over time, yet, as we will see, it is simple enough to work with easily and to provide some analytic estimation results.

3 3-d estimation in a simplified rendering domain

Our goal is to infer 3-dimensional human motion from an image sequence. We first study the problem under simplified rendering conditions, which helps us understand the problem and provides analytic results. These results will give us an upper bound on the estimation accuracy achievable with natural images, and will also motivate how to process the natural scenes.

Our simplified rendering conditions are as follows: the body is transparent, and each marker is rendered to the image plane orthographically. For figural motion described by human motion basis coefficients $\vec{\alpha}$, the rendered image sequence, $\vec{y}$, is:

$$\vec{y} = PU\vec{\alpha}, \tag{1}$$

where P is the projection operator which collapses the y dimension of the image sequence $U\vec{\alpha}$. Note that under these rendering conditions, the markers are distinguishable from each other.

To estimate the figure's 3-d motion, we want to find the most probable 3-d explanation, specified by $\vec{\alpha}$, for a given 2-d observation of markers over time, $\vec{y}$. By Bayes theorem, we have

$$P(\vec{\alpha}|\vec{y}) = k_1 P(\vec{y}|\vec{\alpha})P(\vec{\alpha}), \tag{2}$$

where k_1 is a normalization constant independent of the parameters $\vec{\alpha}$ that we seek to optimize. As developed above, for the prior probability, $P(\vec{\alpha})$, we

have our multi-dimensional gaussian,

$$P(\vec{\alpha}) = k_2 e^{-\vec{\alpha}' \Lambda^{-1} - \vec{\alpha}}, \tag{3}$$

where k_2 is another normalization constant. If we model the observation noise as i.i.d. gaussian with variance σ, we have, for the likelihood term of Bayes theorem,

$$P(\vec{y}|\vec{\alpha}) = k_3 e^{-|\vec{y} - PU\vec{\alpha}|^2/(2\sigma^2)}, \tag{4}$$

with normalization constant k_3.

The posterior distribution is the product of these two gaussians. That yields another gaussian, with mean and covariance found by a matrix generalization of "completing the square" [7]. The squared error optimal estimate for α is then

$$\alpha = SU'P'(PUSU'P' + \sigma I)^{-1}(\vec{y} - (P\vec{m})) \tag{5}$$

.

Figure 5 illustrates applying this estimate to an overlapped sequence of 3 motion segments (20 frames, each 10 frame segment offset by 5 frames). We omitted one of the 10 sequences from the training data, and used a subset of it for this test. (a) shows the original sequence, and (b), the orthographic projection. (c) is the 3-d reconstruction resulting from the likelihood term alone, omitting the gaussian prior information. This finds the coefficients of the human motion basis functions which best explain the visual data. Note that the 3-d reconstruction is poor. (d) is the full Bayesian solution of Eq. 5: including the prior information gives a much better 3-d reconstruction. Our gaussian probability model and the simplified rendering conditions allow this analytic solution for the optimal 3-d motion estimate.

We also know the covariance matrix, Q, describing the uncertainty in the estimated 3-d configuration after viewing the 2-d sequence,

$$Q = S - SU'P'(PUSU'P' + \sigma I)^{-1}PUS, \tag{6}$$

where I is the identity matrix, the rows and columns having the dimensionality of the observations.

Of course, without any human motion model, the depth of each marker would be completely unknown; our prior model for human motion removes most of those ambiguities. The structure of the posterior covariance Q reveals what ambiguities that remain in the 3-d structure after viewing the image

sequence. Figure 3 compares the diagonal terms of the prior covariance (solid line) with those of the posterior covariance (dashed line). One mode, mode 2, shows virtually no reduction in uncertainty; that corresponds to a rigid translation mode moving nearly along the line of sight of the camera, shown in Fig. 6 (a). The second highest uncertainty mode, mode 1, is another rigid translation mode, shown in Fig. 6 (b). The non-rigid mode having the highest posterior uncertainty, mode 10, is shown in Fig. 6 (c). This mode spreads the arms along the line of sight of the camera. We note that this high uncertainty mode reflects the errors observed in the reconstruction of Fig. 5 (d).

Under our gaussian prior model, we can quantify how much 3-d motion information we gain from seeing the orthographic projections of the marker positions over time. For the example of Fig. 5, the prior probability distribution occupies a certain volume in the 50 dimensional parameter space. The ratio of that volume to the posterior probability distribution's volume is 10^{-14}. While it is hard to gauge high dimensional volumes intuitively, the ratio is small; the posterior uncertainty is considerably reduced. In post-processing, we might expect to remove also the rigid mode ambiguities, either by user interaction, or by applying rigid ground contact constraints.

We draw several conclusions from studying the problem in this simplified rendering domain. Using prior knowledge of human motion indeed does improve the 3-d reconstructions possible from monocular image sequence data. For our training set of human motions, the remaining uncertainty after observations lay in the rigid translation away from the camera, and in a mode spreading the arms along the camera ray. The reconstructions are generally good. The image information we have used are the 2-d projections of marker positions of a stick figure. We conclude that if we are able to accurately overlay a 2-d stick figure on top of the human figure in a video sequence, approximately orthrographically rendered, we should be able to achieve comparable 3-d reconstruction accuracy from real images.

4 3-d estimation with real images

In order to estimate the 3-d body motion, we first want to find a stick figure summary of the 2-d moving image of a human figure. This is a problem that various research groups have addressed, and, to a large degree, solved. Hager

and Belheumer, and Black and collaborators have developed parameterized motion models for tracking particular human actions [9, 2]. Blake and collaborators [3] have developed contour-based tracking of non-rigid objects. Pfinder [15] tracks the human figure over stationary environments.

We developed our own tracking method, although code from the tracking methods of other groups should also work. Because we wanted to reconstruct 3-d figures even from difficult film sequences, we allowed ourselves interactive correction of the mistakes of the automatic tracking. This is in addition to the other interactions needed for the problem: to specify which human to track, and over which time period.

Our goal was to demonstrate that the reconstruction method developed for the simplified rendering carried over to natural scenes. We assumed that the image rendering, over the time frames of interest, was roughly orthographic. We ignored the effect of a moving background; we reconstructed the figure's motion relative to the camera frame, not the background.

To achieve some independence of clothes patterns of the human figure, we took the gradient of the image intensities, and normalized the gradient strengths by a blurred average of the local contrast strength. We then blurred these normalized edge strengths enough to make a low-resolution sampling of 10 by 8 sensors.

Based on the location of the sticks of our stick figure, we formed a prediction for what the sensors ought to see, assigning a fixed edge strength to each stick. We penalize the squared difference between the observed sensor responses and the predictions.

The user can interactively specify the correct location of any stick figure part at any time frame. This effectively places a spring between the image position and the stick figure part at that particular time.

These two inputs are integrated with the prior information in an function optimization scheme. We seek the $\vec{\alpha}$ which minimizes an energy, $E(\vec{\alpha})$,

$$E = (\vec{R} - \vec{f}(\vec{\alpha}))^2 + \lambda_1(\vec{\alpha}'\Lambda^{-1}\vec{\alpha}) + \lambda_2 \sum_i (\vec{I}_i - P_i\vec{\alpha})^2. \qquad (7)$$

$\vec{R}$ is the vector of sensor responses over time from the image data. The function $\vec{f}$ converts $\vec{\alpha}$ body motion coefficients to predicted sensor responses. $\vec{I}_i$ is the ith point position specified by the user, and P_i projects the α coefficients onto the corresponding ith stick figure part 2-d position. λ_1 and λ_2 are

constants which reflect the weights of the image data, the priors over human motions, and the interactively specified 2-d point matches.

In the Bayesian framework, we interpret E as the negative log of the posterior probability. λ_1 and λ_2 then represent observation and user "noise strengths". The quadratic penalty for sensor response differences is the log of the likelihood term, and both the interactively placed springs and the gaussian prior motion model represent prior information about the parameters. (We also included a 90° rotated version of all our training data in the calculation of the prior probability.) We find the 3-d body motion parameters which maximize the posterior probability. This code runs in interactive time in Matlab on an SGI Onyx.

The recovered optimal $\vec{\alpha}$ yields the recovered marker positions over time. We then fit those marker positions to cylinder positions in a simple figure model using least squares techniques.

Figure 7 shows the resulting estimates of the 3-d figure positions from a 100 frame sequence of Barishnikov dancing. In order to test our 3-d reconstruction algorithm, rather than our 2-d tracking algorithm, we used approximately one interactive location specification per frame, to ensure the accurate overlay of a stick figure over the motion sequence. Figure 7 (a) shows the input sequence and the overlaid stick figure.

We minimized E in Eq. 7 to find the $\vec{\alpha}$ estimate over each overlapped 10-frame segment. Positions from the overlapped segments were linearly blended together. We set the offset away from the camera of each segment (the rigid mode we can't estimate) to ensure continuity to the next segment.

Figure 7 (b) shows the recovered 3-d marker positions, viewed from 30° away from the camera position. Figure 7 (c) shows the 3-d cylinder model, viewed from that same off-camera position. Given the simple gaussian prior model for human motion, we feel the results are remarkably strong. We have rendered the 3-d dance as a linear combination of basis motions learned from our motion capture training set. The dancing cylinder figure generally captures the 3-d motions of the dancer. The cylinder dancer does not point his toes, as Barkshnikov does, but toe pointing was not in the training set.

5 Discussion

Our approach has no notions of balance, support, or friction, apart from what was learned from the training data. This suggests it may blend well with an approach based more an physical models of a human (e.g., [11]). Such models may offer realistic physical constraints, but may be difficult to fit from image data, while our approach offers the complementary advantages.

A stronger prior should give our method better results. Despite the analytic appeal of a single gaussian model, a more complex model, such as a mixture of gaussians [1], may model human motions better, and give better 3-d reconstructions.

6 Summary

We have used a statistical model to infer the 3-d positions of a human figure from an image sequence of the human moving. We learned our model of human motion from a training set of 3-d motion capture data, obtained from a calibrated tracking system.

We found it fruitful to first explore the 3-d recovery problem in a simplified rendering model (markers on a transparent figure, viewed under orthography). This linear rendering yields analytic solutions for the mean-squared optimal 3-d motion estimate, as well as a covariance estimate for the posterior uncertainty in the human motion. This identifies the rigid and non-rigid motion modes that are most difficult to estimate from the 2-d motion sequence.

Those results show how to estimate the 3-d figure motion if we can place a 2-d stick figure over the image of the moving person. We developed such a tracker, allowing interactive correction of tracking mistakes, to test our 3-d recovery method. We show good recovery of 3-d motion for a difficult dance sequence, viewed from a single camera. These results show the power of adding prior knowledge about human motions, in a Bayesian framework, to the problem of interpreting images of people.

References

[1] C. M. Bishop. *Neural networks for pattern recognition.* Oxford, 1995.

[2] M. J. Black, Y. Yacoob, A. D. Jepson, and D. J. Fleet. Learning paramterized models of image motion. In *Proc. IEEE CVPR*, pages 561–567, 1997.

[3] A. Blake and M. Isard. 3D position, attitude and shape input using video tracking of hands and lips. In *Proc. SIGGRAPH 94*, pages 185–192, 1994. In *Computer Graphics*, Annual Conference Series.

[4] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley-Interscience, 1973.

[5] I. Essa, editor. *International Workshop on Automatic Face- and Gesture- Recognition.* IEEE Computer Society, Killington, Vermont, 1997.

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

[7] A. Gelb, editor. *Applied optimal estimation.* MIT Press, 1974.

[8] L. Goncalves, E. D. Bernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3d. In *Proc. 5th Intl. Conf. on Computer Vision*, pages 764–770. IEEE, 1995.

[9] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. In *Proc. IEEE CVPR*, pages 403–410, 1996.

[10] I. A. Kakadiaris and D. Metaxas. 3d human body model acquisition from multiple views. In *Procp. 5th Intl. Conf. on Computer Vision*, pages 618–623. IEEE, 1995.

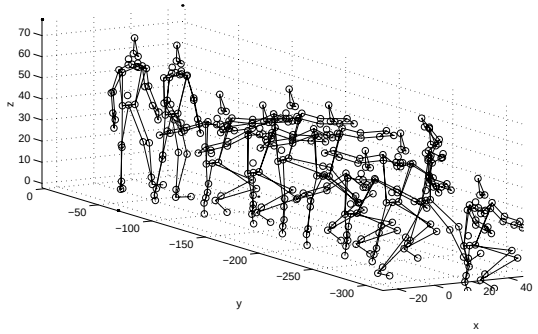
[11] H. Ko and N. Badler. Animating human locomotion in real-time using inverse dynamics, balance and comfort control. *IEEE Computer Graphics and Applications*, 16(2):50–59, 1996.

[12] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *Proc. 5th Intl. Conf. Computer Vision*, pages 786–793. IEEE, 1995.

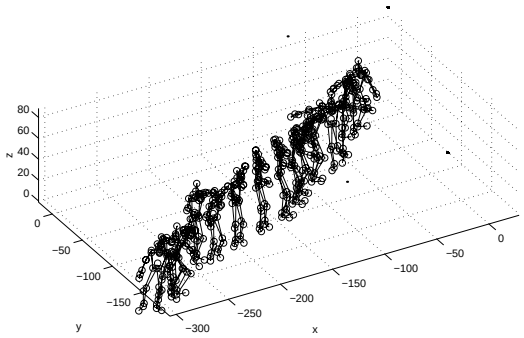
[13] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. 5th Intl. Conf. on Computer Vision*, pages 612–617. IEEE, 1995.

[14] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1), 1991.

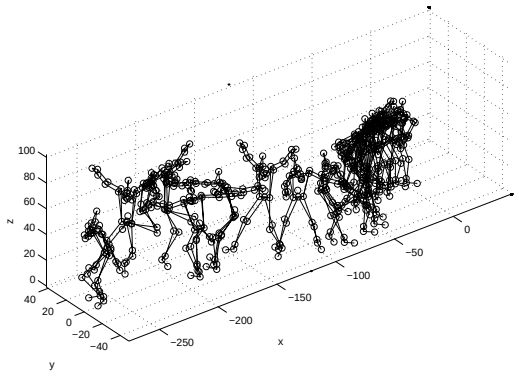
[15] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. In I. Essa, editor, *Intl. Conf. on Automatic Face and Gesture Recognition*, pages 51–56, Killington, Vermont, 1996. IEEE Computer Society.



(a)

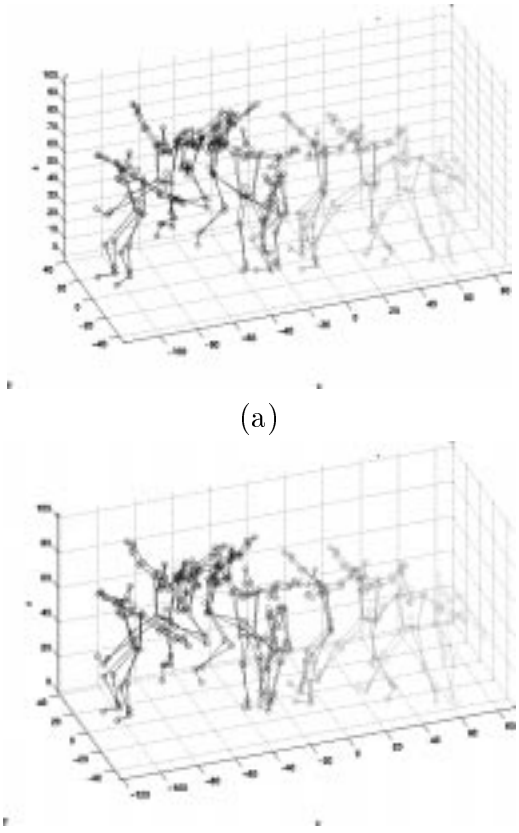


(b)



(c)

Figure 1: Example motion sequences from our training set of 10 sequences.



(a) (b)

Figure 2: Figure showing approximation of the human motion signal as a linear combination of basis functions. (a) 40 basis functions approximation; (b) original.

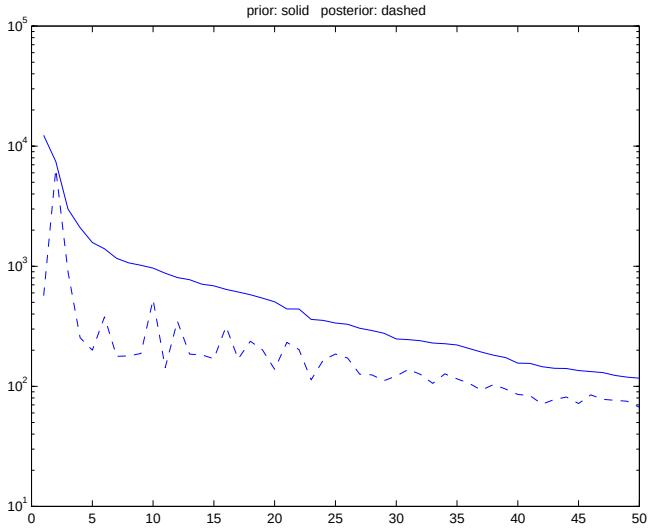
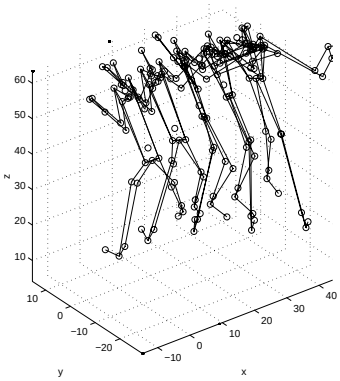
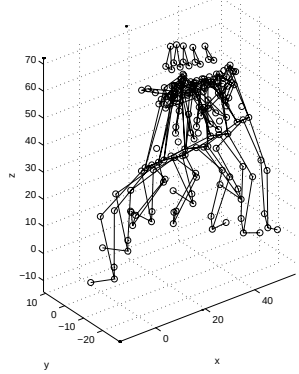


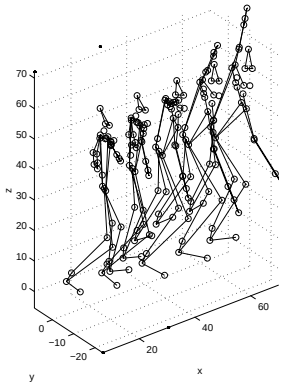
Figure 3: Prior (solid line) and posterior (dotted line) covariance matrix diagonal elements, plotted on a semi-log scale. Measurment of the marker positions in the image plane dramatically reduces the prior uncertainty of the human motion. The one mode where uncertainty is not reduced corresponds to rigid motion along the line of sight.



(a)



(b)



(c)

Figure 4: Three random draws from the gaussian prior distribution over the 37 3-d marker positions. Note that they all look human, and correspond roughly to human motions

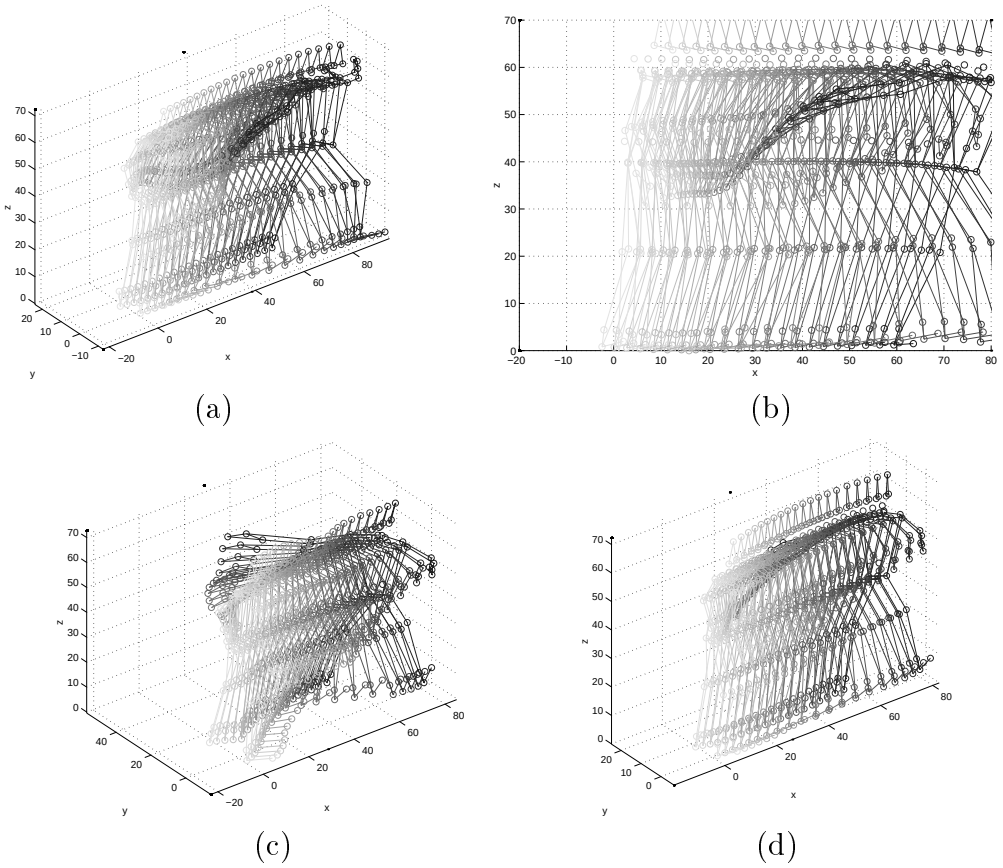
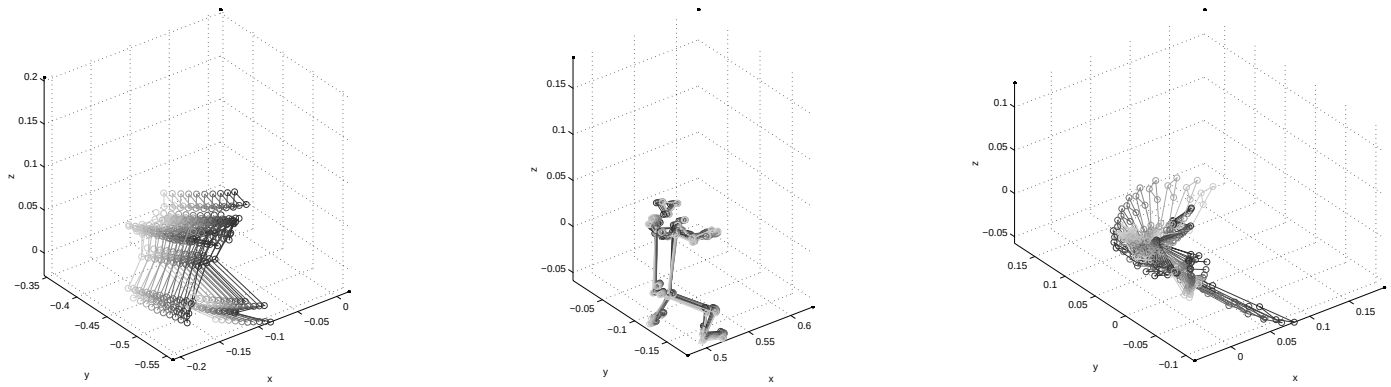


Figure 5: (a) Original 3-d sequence. (b) Orthographically projected image (markers connected by lines for clarity). (c) 3-d reconstruction omitting prior information. This is the 3-d figure in the eigenspace of the human motions which best accounts for the image data. (d) Full Bayesian reconstruction. Note that the addition of prior information creates a reconstruction more similar to the original. The high posterior covariance modes (Fig. 6) explain the remaining differences from the original sequence.



(a) (b) (c)

Figure 6: The three modes with the highest posterior uncertainty.

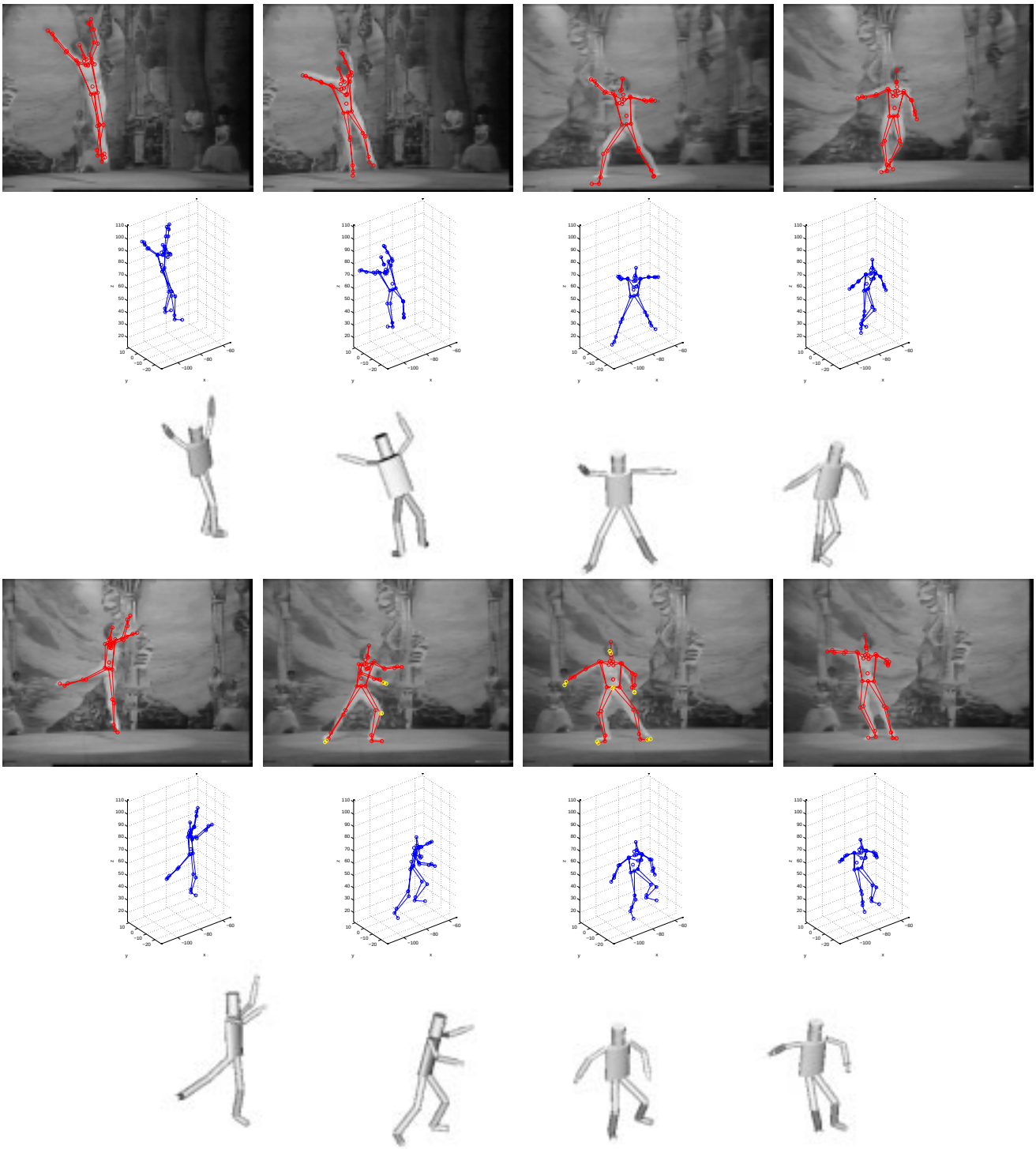


Figure 7: Top rows: Samples from 100 frame sequence of Barishnikov dancing, with tracked 2-d stick figure overlaid. Middle: Inferred 3-d (marker) positions, using gaussian prior model for 3-d figure motions. Bottom: Recovered 3-d moving cylinder figure, which generally captures well the 3-d motions of the dancer. (Please see *ftp* site movie.)