

# Estimating 3D Body Pose using Uncalibrated Cameras

Rómer Rosales, Matheen Siddiqui, Jonathan Alon, and Stan Sclaroff  
Computer Science Department, Boston University  
Boston, MA 02215

## Abstract

*An approach for estimating 3D body pose from multiple, uncalibrated views is proposed. First, a mapping from image features to 2D body joint locations is computed using a statistical framework that yields a set of several body pose hypotheses. The concept of a “virtual camera” is introduced that makes this mapping invariant to translation, image-plane rotation, and scaling of the input. As a consequence, the calibration matrices (intrinsic) of the virtual cameras can be considered completely known, and their poses are known up to a single angular displacement parameter. Given pose hypotheses obtained in the multiple virtual camera views, the recovery of 3D body pose and camera relative orientations is formulated as a stochastic optimization problem. An Expectation-Maximization algorithm is derived that can obtain the locally most likely (self-consistent) combination of body pose hypotheses. Performance of the approach is evaluated with synthetic sequences as well as real video sequences of human motion.*

## 1. Introduction

The estimation of 3D human body structure from visual cues is a key problem faced by computer vision. Recovery of detailed body pose from images would enable a great number of applications, including human-computer interfaces, video coding, visual surveillance, human motion recognition, ergonomics, video indexing/retrieval, etc.

This paper introduces a framework for 3D articulated pose recovery, given multiple uncalibrated views. The map from visual features to body joint locations is obtained via a statistical inference method, known as Specialized Mappings Architecture (SMA) [20, 19]. The SMA provides several pose hypotheses, each one with correspondences of 2D joint locations across frames and views. From the set of pose hypotheses, 3D pose can be recovered via multiple-view geometry and an alternating minimization algorithm.

One strength of our approach is due to the fact that the camera matrices used in the 3D pose recovery are not those of the actual cameras that captured the sequence, but rather the *virtual cameras* with which the SMA was trained. The calibration matrices of these virtual cameras can be regarded as completely known; furthermore, their pose is known up to a single angular displacement parameter. In other words, no camera calibration is required, and there is only one parameter per camera (excluding the first camera)

to recover. Another strength is that our formulation provides a principled way to combine multiple pose hypotheses in a probabilistic form.

## 2. Related Work

Humans can easily estimate body part location and 3D structure from relatively low-resolution images of the projected 3D world (e.g., watching a video). Unfortunately, this problem is inherently difficult for a computer. The difficulty stems from the number of degrees of freedom in the human body, ambiguities in the projection of human motion onto the image plane, self-occlusion, insufficient temporal or spatial resolution, etc. To make the problem tractable, many 3D pose estimation algorithms use human body models, and/or prior knowledge obtained via machine learning.

Three-dimensional models have been used in systems that estimate and track body pose in image sequences [3, 6, 8, 15, 17, 18, 21, 25]. Unfortunately, most of these tracking methods require careful initialization of the 3D model on the first frame, and tracking in subsequent frames tends to be sensitive to errors in initialization, and numerical drift. In some approaches, it is also assumed that tracking of joint locations (or correspondence of body features) is given in each input image. Three-dimensional models have also been employed in estimation of human body pose from a single image [1, 14, 23]; however, most of these methods also require that projected joint locations be given as input.

Recently, researchers proposed the use of machine learning methods that exploit prior knowledge in gaining more stable estimates of 3D human body pose [2, 9, 22]. In [9], a Gaussian probability model was used for short human motion sequences. In similar approach [2], the manifold of human body configurations was modeled via a hidden Markov model and learned via entropy minimization. In [22] dynamic programming was used to calculate the best global labeling of the joint probability density function of the position and velocity of body features. In these systems, the joint locations, correspondences, and/or model initialization must be provided by hand. Moreover, when **general** human motion dynamics is intended to be learned (besides just configurations), the requirements of amount of data, model complexity, and computational resources, become impractical. As a result, models with large priors towards specific motions, like walking motions, are generated.

In [19, 20] a machine learning method, the Specialized

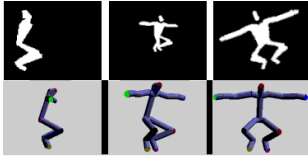


Figure 1: Silhouettes segmented from real camera views, and 2D joint estimates in the corresponding virtual camera views.

Mapping Architecture (SMA) was proposed. Our work builds upon the results presented here. The SMA framework allows mapping directly from image features to 2D image locations of body joints. The SMA’s mapping functions are estimated from training data, in this case: examples of body poses (output) and their corresponding visual features (input). The SMA is related to other machine learning models [7, 11] and mixture models in general that use the principle of divide-and-conquer to reduce the complexity of the learning problem.

### 3 The Basic Idea of Our Approach

In this paper, a SMA is used to compute a mapping from image features to corresponding 2D joint locations in the image planes of *virtual cameras*. These virtual camera views are a direct consequence of our use image features (Hu moments [10]) that are invariant to translation, scaling, and rotation on the image plane. This is evident from Fig.1. Note that these estimates are insensitive to image translation, scaling and rotation. Also note the overall vertical orientation of the skeleton. This is due to the fact that the SMA mappings were trained with concentric cameras, where each camera’s principal axis passes through the circles’ center, and all cameras have coinciding up-vectors; in other words, the virtual image scan lines should ideally match across views.

Given correspondences of the most likely 2D joint locations in multiple, virtual camera views, obtained from the SMA inference procedure, 3D body pose can be recovered with an algorithm related to structure from motion. Moreover, given multiple hypothesis per camera (as in this paper) we formalize a generalized probabilistic structure from motion technique and provide an algorithm for the special virtual camera case.

An overview of our approach is shown in Figs. 2–3. Our goal is to produce a reconstruction of body pose  $\mathbf{X}$  (3D joint locations) and the relative orientations  $\Omega$  of the *real* cameras. First images obtained by each camera  $c$  are segmented to extract image features  $\mathbf{x}_c$  (Hu moments in this work). Each camera is assumed to capture images of the whole body, camera parameters are otherwise unconstrained.

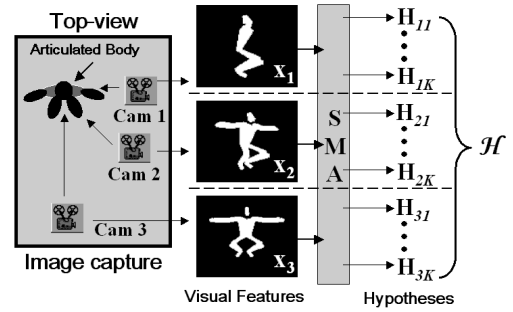


Figure 2: Overview of pose hypothesis generation. Cameras capture the scene, human figures are segmented, and a vector of visual features  $\mathbf{x}_c$  is computed per camera. SMA produces several body pose hypotheses per camera.

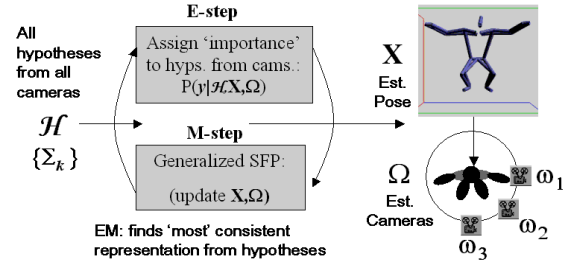


Figure 3: Overview of structure from motion algorithm. All body pose hypotheses and error covariances are used to find a self-consistent combination of hypotheses using EM. The output are the estimates of body pose and camera parameters.

#### 3.1 Visual Features to 2D Pose

The front-end of our approach consists of a Specialized Mapping Architecture (SMA) [20, 19]. For the purpose of this work, an SMA may be regarded as a set of functions  $\Phi = \{\phi_k\}$  that have been trained to map inputs to outputs. The SMA mapping functions have been precomputed via a supervised learning procedure. Training data for the SMA is generated via computer graphics renderings of 3D motion capture data as detailed in [20, 19]. As shown in Fig. 2, each function  $\phi_k$  transforms (maps) a vector of visual features  $\mathbf{x}$  into several 2D body pose hypotheses,  $\mathbf{H}_k$ . Each  $\mathbf{H}_k$  gives an hypothesis about the image locations of the body joints for that particular view.

Given a silhouette for the human extracted in  $C$  camera views, each yielding feature vector  $\mathbf{x}_c$ , we denote  $\mathbf{H}_{c,k}$  as the hypothesis  $k$  from camera  $c$ . Note that each camera uses the same series of functions  $\Phi$  to produce its hypothesis, thus  $\Phi$  are not trained for a particular camera viewpoint. We denote the set of all hypotheses from all cameras  $\mathcal{H}$ . Fig. 4 shows an example of what  $\mathcal{H}$  looks like for  $C = 3$  cameras and  $K = 4$  hypothesis. This is a real example, where input features were obtained from the images shown in Fig. 1. The vertical bars indicate which hypothesis is deemed most likely by the SMA for each view.

### 3.2 3D Structure and Camera Estimation from Multiple Pose Hypotheses

The individual *best* configurations shown in Fig. 4 (marked with a bar) are not guaranteed to be self-consistent, i.e., they may not agree. A 3D reconstruction algorithm needs to account for this inconsistency of the observations. In response to this, 3D estimation is posed as a maximum likelihood (ML) estimation problem that tries to find the best estimates for  $\mathbf{X}$  and  $\Omega$  given the set of all hypotheses from all cameras  $\mathcal{H} = \{\mathbf{H}_{c,k}\}$ . As shown in Fig. 3 we try to find a 3D body pose and cameras that generate the most consistent combination of 2D hypotheses from each camera.

In the following sections we formulate this problem and show its exact solution is intractable. We then present an Expectation Maximization algorithm that approximates a solution to the initial problem. In order to prepare the reader for the next sections, we will anticipate that the form of the EM algorithm requires solving a *generalized* form of the *Structure from Motion* problem at each iteration of the M-step. This is related in some sense to the recent work of [4], but differs in the following key aspects: 1.) our method handles multiple hypotheses per camera, 2.) our method uses a full covariance matrix *weighting* for the hypotheses in the solution, and 3.) sampling methods are not required because our problem is posed in a tractable form.

## 4 Probabilistic 3D Reconstruction

Reconstruction will be formulated as a stochastic optimization problem. The goal is to maximize the log-likelihood of the 3D body pose  $\mathbf{X}$  and camera parameters  $\Omega$ :

$$\Omega^*, \mathbf{X}^* = \arg \max_{\mathbf{X}, \Omega} \log p(\mathcal{H} | \mathbf{X}, \Omega), \quad (1)$$

where  $\mathbf{X} \in \mathbb{R}^p$ , with  $p = 3 \times (\text{number of joints})$  and  $\Omega = (\omega_1, \dots, \omega_C)$ . Assuming that the hypotheses presented by each camera are conditionally independent given the model parameters  $\mathbf{X}$  and  $\Omega$ , we have:

$$\Omega^*, \mathbf{X}^* = \arg \max_{\mathbf{X}, \Omega} \log \prod_{c=1}^C p(\mathbf{H}_c | \mathbf{X}, \Omega). \quad (2)$$

this assumption is reasonable because by knowing the true value of  $\mathbf{X}$  and  $\Omega$ , we do not gain any information about a given view if we know about another view. Introducing a latent random variable  $y_c \in \{1..K\}$  representing the choice of hypothesis for camera  $c$  we obtain:

$$\Omega^*, \mathbf{X}^* = \arg \max_{\mathbf{X}, \Omega} \sum_{c=1}^C \log \left( \sum_{k=1}^{K_c} p(\mathbf{H}_c | y_c = k, \mathbf{X}, \Omega) P(y_c = k | \mathbf{X}, \Omega) \right). \quad (3)$$

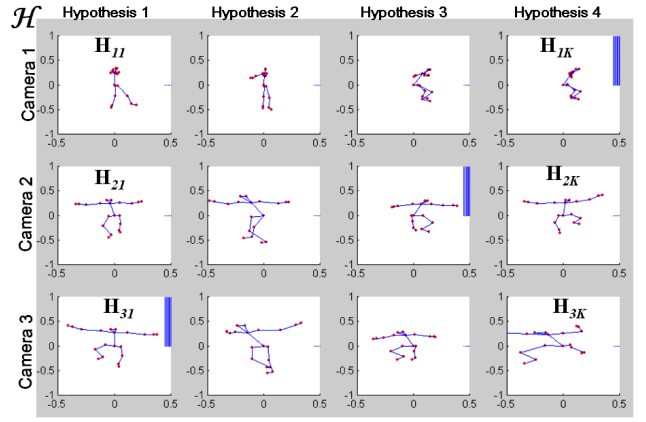


Figure 4: Four hypotheses generated for three cameras views of the same body pose shown in Fig.2. The most likely pose per camera is marked with a vertical bar.

Intuitively,  $p(\mathbf{H}_c | y_c = k, \mathbf{X}, \Omega)$  is how probable the  $k$ -th hypothesis of camera  $c$  is, given the model parameters  $\mathbf{X}, \Omega$ . It is important to recall that the probability of a set of hypothesis from several cameras can be factorized into a product of probabilities of hypotheses from each camera (because of our conditional independence assumption). Thus, combinatorial complexity is avoided.

If we consider  $P(y_c = k | \mathbf{X}, \Omega)$  as uniform, then we face the following stochastic optimization problem:

$$\Omega^*, \mathbf{X}^* = \arg \max_{\mathbf{X}, \Omega} \sum_{c=1}^C \log \left( \sum_{k=1}^{K_c} p(\mathbf{H}_c | y_c = k, \mathbf{X}, \Omega) \right). \quad (4)$$

Because of the log-sum encountered, this problem is intractable in general. However, there exist practical approximate optimization procedures, one of them is Expectation Maximization (EM) [5, 13].

## 5 EM Algorithm for Estimating 3D Body Pose and Virtual Cameras

We will now present the EM algorithm parameter update rules for the specific problem at hand.

The E-step consists of finding  $P(y = k | \mathbf{H}, \mathbf{X}, \Omega) = \tilde{P}(y)$ . Note that the variables  $y_c$  are independent (it follows from our conditional independence assumption in Sec. 4). Therefore, we can factorize  $\tilde{P}(y) = \prod_c \tilde{P}(y_c)$ . Assuming a uniform prior over the specialized functions of any camera, i.e.,  $P(y_c = k | \mathbf{X}, \Omega) = \alpha$ , it can be shown that:

$$\tilde{P}(y_c) = p(\mathbf{H}_c | y_c = k, \mathbf{X}, \Omega) / \sum_j p(\mathbf{H}_c | y_c = j, \mathbf{X}, \Omega)$$

However, note that  $p(\mathbf{H}_c | y_c = k, \mathbf{X}, \Omega)$  is still undefined. In this paper we use:

$$p(\mathbf{H}_c | y_c = k, \mathbf{X}, \Omega) = \mathcal{N}(\mathbf{H}_{c,k}; R(\mathbf{X}, \Omega_c), \Sigma_k), \quad (6)$$

with  $\mathbf{H}_{c,k}$  the  $k$ -th hypothesis of camera  $c$ ,  $\Omega_c$  the parameters of the camera  $c$ , and  $\Sigma_k$  the covariance error of the specialized function  $k$  (see [19] for a detailed treatment of specialized maps).

One way to interpret this choice is to simply think that the error cost in the projection of the current estimate is a Gaussian distribution. This seems a natural choice, and leads to tractable further derivations. The distribution is not spherical, but shaped according to  $\Sigma_k$  to represent our degree of confidence in the hypotheses generated by the SMA.

## 5.1 M-step

The M-step consists of finding

$$(\Omega, \mathbf{X})^{(t)} = \arg \max_{\Omega, \mathbf{X}} E_{\tilde{P}^{(t)}} [\log p(\mathbf{H}, \mathbf{y} | \Omega, \mathbf{X})]. \quad (7)$$

In our case we can show that this is equivalent to:

$$\begin{aligned} \arg \max_{\Omega, \mathbf{X}} \sum_c E_{\tilde{P}^{(t)}} [\log p(\mathbf{H}_{c,k}, y_c | \Omega, \mathbf{X})] = \\ \arg \min_{\Omega, \mathbf{X}} \sum_c \sum_k \tilde{P}^{(t)}(y_c = k) \\ (\mathbf{H}_{c,k} - R(\mathbf{X}, \Omega_c))^T \Sigma_k^{-1} (\mathbf{H}_{c,k} - R(\mathbf{X}, \Omega_c)) \end{aligned} \quad (8)$$

Very interestingly, Eq. 8 corresponds to a generalized version of the *Structure from Motion* problem [16]. In the standard structure from motion problems,  $\Sigma_k$  is assumed diagonal and there is only one hypothesis per camera (i.e., there is only one observation). Here, there are several hypotheses per camera  $\mathbf{H}_{c,k}$  with  $k = 1..K$ , provided by the specialized maps [20]. Thus, our formulation generalizes the structure from the motion problem and provides a probabilistic framework for its solution.

If the rendering function  $R$  merely projects the 3D joints to the different virtual cameras views then the resulting cost function to minimize can be written as:

$$\begin{aligned} J(\Omega, \mathbf{X}) = \sum_c \sum_k \tilde{P}^{(t)}(y_c = k) \\ (\mathbf{H}_{c,k} - M(\Omega_c)\mathbf{X})^T \Sigma_k^{-1} (\mathbf{H}_{c,k} - M(\Omega_c)\mathbf{X}) \end{aligned} \quad (9)$$

where  $\mathbf{H}_{c,k}$  is a  $2N$  vector ( $N$  is the number of joints) of the 2D joints estimates,  $\mathbf{X}$  is a  $3N$  vector of the 3D joints estimates,  $\Sigma_k$  is a  $2N \times 2N$  covariance matrix of the 2D joints estimates, and  $M$  is a  $2N \times 3N$  matrix consisting of  $N$  copies of a  $2 \times 3$  affine camera matrix along the diagonal.

The cost function  $J$  is in general nonlinear in its parameters; however, the partial derivatives of  $J$  are bilinear in both the structure ( $\mathbf{X}$ ) and camera ( $M_c$ ) parameters when the affine projection model is assumed [12]:

$$\frac{\partial J}{\partial \mathbf{X}} = \sum_c \sum_k \tilde{P}^{(t)}(y_c = k) M_c^T \Sigma_k^{-1} (\mathbf{H}_{c,k} - M_c \mathbf{X}) \quad (10)$$

$$\frac{\partial J}{\partial \mathcal{M}_c} = \sum_k \tilde{P}^{(t)}(y_c = k) \mathcal{X}^T \Sigma_k^{-1} (\mathbf{H}_{c,k} - \mathcal{X} \mathcal{M}_c). \quad (11)$$

Here,  $\mathcal{X}$  is a  $2N \times 6$  matrix, with the  $i^{th}$   $2 \times 6$  block taking the form:

$$\begin{bmatrix} \mathbf{X}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_i \end{bmatrix}, \quad (12)$$

and  $\mathcal{M}_c$  is a 6D vector consisting of the elements of the  $2 \times 3$  affine camera matrix in row major order.

By setting the partials to zero and rearranging the resulting equations, we obtain a set of bilinear equations for structure  $\mathbf{X}$  and camera  $\mathcal{M}_c$ :

$$\begin{aligned} \left( \sum_c \sum_k \tilde{P}^{(t)}(y_c = k) M_c^T \Sigma_k^{-1} M_c \right) \mathbf{X} = \\ \sum_c \sum_k \tilde{P}^{(t)}(y_c = k) M_c^T \Sigma_k^{-1} \mathbf{H}_{c,k} \end{aligned} \quad (13)$$

and

$$\begin{aligned} \left( \sum_k \tilde{P}^{(t)}(y_c = k) \mathcal{X}^T \Sigma_k^{-1} \mathcal{X} \right) \mathcal{M}_c = \\ \sum_k \tilde{P}^{(t)}(y_c = k) \mathcal{X}^T \Sigma_k^{-1} \mathbf{H}_{c,k} \end{aligned} \quad (14)$$

Note that Eq. 14 is a set of six linear equations in the camera's affine parameters; however, in our case, the affine (virtual) camera model is simple, and has only one degree of freedom. The orientation of the virtual camera is:

$$\begin{aligned} \mathcal{M}_c &= [\cos \omega_c \quad \sin \omega_c \quad 0 \quad 0 \quad 0 \quad 1]^T \\ &= [a \quad b \quad 0 \quad 0 \quad 0 \quad 1]^T \end{aligned} \quad (15)$$

Therefore, we solve the above over-constrained set of linear equations to obtain  $a$  and  $b$ , and then enforce the nonlinear constraint  $\omega_c = \tan^{-1}(\frac{b}{a})$ .

To solve, we start with an initial guess for the camera parameters  $M_c$ , and use Eq. 13 to obtain the least squares solution for  $\mathbf{X}$ . Then this new  $\mathbf{X}$  and Eq. (14) are used to solve for the camera parameters  $\mathcal{M}_c$ . This step is repeated until convergence is achieved. The initial guess for the camera parameters  $M_c$  is obtained from a modified version of the standard factorization algorithm [24], which does not incorporate the  $y$  locations in the measurement matrix. Details are given in the Appendix.

## 5.2 Multiple Frames

For the sake of simplicity, our formulation was derived for a single frame only. In our implementation, we extended it to multiple frames. A straightforward way to do this is to assume that frames are conditionally independent over time:  $p(\mathcal{H}_1, \dots, \mathcal{H}_T | \mathbf{X}_1, \dots, \mathbf{X}_T, \Omega) = \prod_{t=1}^T p(\mathcal{H}_t | \mathbf{X}_t, \Omega)$ , with  $t$  indicating the frame number.

This is different than just several single frame estimations in that  $\Omega$  is not time dependent, and therefore more evidence from multi-frames should tend in theory to provide a more robust estimate. An equivalent to Eqs.13 and 14

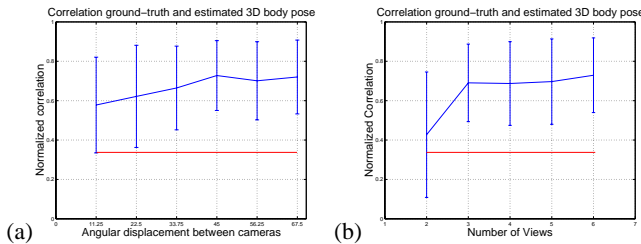


Figure 5: Normalized correlation between ground-truth and estimated 3D poses: (a) as a function of the angular displacement between cameras (using three cameras), and (b) as a function of the number of camera views employed. For comparison, the mean normalized correlation obtained for 100 randomly chosen pairs of 3D poses in our test set was 0.34.

was derived with just an additional summation over time  $t$  in each side. Note also that these total independence assumptions can be relaxed easily to obtain, for example, Markov-like models of a given order. In these cases, inference is not as efficient and close approximations may be harder to find.

## 6. Experiments

The training data for the SMA consists of approximately 35 sequences (5,000 frames) obtained from 3D motion capture data (including basic locomotion, dancing, grabbing, throwing, jumping, signaling, and crouching-down). Input-output pairs for training (and testing) were generated using computer graphics by rendering from 32 viewpoints uniformly sampled on the sphere equator. The input consisted of seven real-valued Hu moments [10] computed on synthetically generated silhouettes. The output of the SMA consisted of 20 joint locations (40 DOF) linearly encoded by nine real values using Principal Component Analysis (PCA).

Experiments have been conducted using real video sequences, and synthetic sequences that were not used in training the SMA. For the experiments with real video sequences, observation inputs were obtained automatically using simple background subtraction. Approximately 105,000 2D configurations were generated synthetically; 8,000 were used for training the SMA and the rest for testing. In the following three performance experiments, unless otherwise stated, three cameras are used to capture the test sequences.

### 6.1 Quantitative Experiments

The first experiment tested the sensitivity of the system to the change in the angular displacement between the cameras, or equivalently the baseline between pairs of cameras. The result of 100 reconstruction trials with 50 randomly chosen frames each is depicted in Fig. 5(a). As seen in the graph, the normalized correlation between ground-truth and estimated 3D pose improves as the baseline increases. A

wider baseline in other approaches results in a major feature correspondence problem. Our approach does not suffer from this shortcoming because the mapping between silhouette features and the 2D joint locations provides correspondence across frames and disparate views.

The second experiment measured the average performance of our approach with respect to the number of cameras used. The angular displacement was set to  $\pi/8$  rads (22.5 degrees) and the number of cameras varied from two to six. Number of trials and performance measures are as in the first experiment. A graph showing results of this experiment is shown in Fig. 5(b). It was noted that there is a major increase in reconstruction accuracy when the number of cameras increases from two to three.

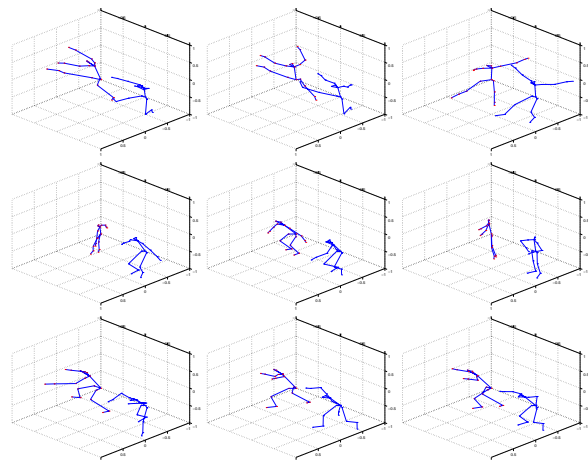


Figure 6: Example reconstruction of a test sequence (every 50-th frame at 30 frames/sec). In each 3D graph, pose estimates are drawn on the left side and ground-truth one the right.

Fig. 6 shows several example reconstructions taken from our test sequences. Frames were taken evenly spaced (every 50-th at 30 frames/sec) from a novel motion capture sequence. Estimates and ground-truth are on the left and right hand side respectively. The visual agreement between estimates and ground-truth can be easily observed.

### 6.2 Experiments with Real Sequences

Real sequences of a single human subject were captured by a setup of three synchronized cameras for which neither calibration nor pose was known. In Fig. 7 we can see the result of standard background subtraction applied to the original images to obtain the initial silhouettes shown (only those obtained by the first camera are shown). Seven Hu moments were computed for each silhouette and then the SMA inference procedure described in this paper was carried out to obtain estimates of 2D joint locations, and then 3D reconstruction was obtained as shown in the bottom rows of Fig. 7. The figure shows every 15-th frame from the sequence. A similar procedure was used to obtain the results

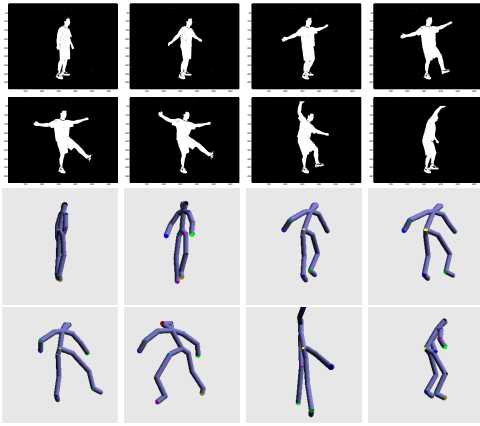


Figure 7: 3D Reconstruction using real video. The top rows show the input images obtained by the first camera. The bottom rows show 3D reconstructions (as viewed by the first camera).

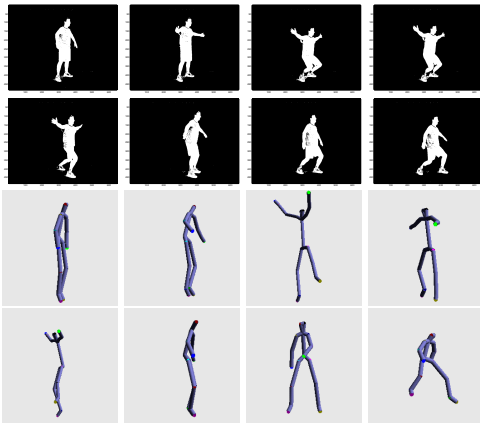


Figure 8: Second example of 3D reconstruction using real video.

in Fig. 8. Cameras were located at roughly mid-body height and their relative angular displacement was approximately 0, 30, and 90 degrees.

We use 50 frames at a time to obtain the poses shown. Estimates of camera angular displacements were 0, 25, and 83 degrees on average. In these sequences, pose accuracy varies over frames, but rough body pose is for the most part visually accurate. Given the complexity of the body configurations, the lack of knowledge about the cameras, and no manual initial placement of the human body model, this is a very difficult task. The main source of inaccuracy when using real data are the statistical differences between the visual features generated by a real person and the computer graphics model used in training the SMA.

## 7. Conclusion and Future Work

This paper has presented a novel approach for estimating 3D body pose from image sequences. We introduced the notion of “virtual cameras” and the formulation of a probabilistic, multiple hypothesis *Structure from Motion* frame-

work along with an approximate, algorithm for solving it.

There are several advantages to the proposed approach. First, it does not require any camera calibration nor manual initialization as required by previous approaches (e.g., [8, 3]) that use multiple-camera setups. Second, the approach exploits the invariances of Hu moments and the concept of “virtual cameras” to obtain good 3D structure estimates and a more tractable algorithm for solving this problem. Third, the approach allows use of large baselines for better 3D reconstruction, without the nettlesome issue of feature correspondence. Finally, the method runs at around 2 frames/sec. when computing with one frame at a time.

Our approach provides approximate estimates of 3D body pose. In these conditions obtaining high accuracy estimates of human body pose is a very ambitious problem in computer vision. Our estimates can be improved by employing kinematic and dynamic constraints, as initial experiments suggest (not included in this paper). Note that this work does not address the image segmentation problem, segmentation is assumed to be “reasonably” good, and is a hard problem on its own.

Even though in theory we could use SMA’s to map directly to 3D pose, the increased dimensionality of the problem makes the SMA training problem computationally less accurate and time consuming and the amount of data required for training also increases. In this paper, we explore the idea that a more complex SMA model might not be required for 3D estimation if camera geometry is incorporated in the solution.

Future work will focus on the derivation of task-specific invariant visual features to better localize 2D joints and the use of more sophisticated kinematic and dynamic priors. Even though we have not exploited these important aspects in this paper, results are encouraging, and indicative of the promise of the overall approach.

## References

- [1] C. Barron and I. Kakadiaris. Estimating anthropometry and pose from a single image. *CVPR*, pp I:669–676, 2000.
- [2] M. Brand. Shadow puppetry. *Proc. ICCV*, pp 1237–1244, 1999.
- [3] C. Bregler. Tracking people with twists and exponential maps. *Proc. CVPR*, 1998.
- [4] F. Dellaert, S.M. Seitz, C.E. Thorpe, and S. Thrun. Structure from motion without correspondence. *Proc. CVPR*, pp II:557–564, 2000.
- [5] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data. *Journal of the Royal Statistical Society (B)*, B 39:1–38, 1977.
- [6] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. *CVPR*, pp II:126–133, 2000.
- [7] J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19,1-141, 1991.
- [8] D.M. Gavrilu and L.S. Davis. 3D model-based tracking of humans in action: A multi-view approach. *Proc. CVPR*, pp 73–80, 1996.

- [9] N. Howe, M. Leventon, and B. Freeman. Bayesian reconstruction of 3D human motion from single-camera video. *Proc. NIPS-12*, 1999.
- [10] M. K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. Inform. Theory*, IT(8), 1962.
- [11] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6, 181-214, 1994.
- [12] D. Morris and T. Kanade. A unified factorization algorithm for points, line segments and planes with uncertainty models. *Proc. ICCV*, pp 696-702, 1998.
- [13] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, M. Jordan (editor), 1998.
- [14] J. O'Rourke and N.I. Badler. Model-based image analysis of human motion using constraint propagation. *PAMI*, 2(6):522-536, Nov 1980.
- [15] A. Pentland and B. Horowitz. Recovery of non-rigid motion and structure. *PAMI*, 13(7):730-742, Jul 1991.
- [16] Hartley R. and Zisserman A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
- [17] J.M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. *Proc. ICCV*, pp 612-617, 1995.
- [18] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP:IU*, 59(1):94-115, 1994.
- [19] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff. 3D hand pose estimation using specialized mappings. *Proc. ICCV*, pp I: 378-385, 2001.
- [20] R. Rosales and S. Sclaroff. Learning body pose using specialized maps. *To Appear Proc. NIPS-14*, 2001.
- [21] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *ECCV*, 2000.
- [22] Y. Song, X. Feng, and P. Perona. Towards detection of human motion. *Proc. CVPR*, pp I:810-817, 2000.
- [23] C.J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Proc. CVPR*, pp I:677-684, 2000.
- [24] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *IJCV*, 9(2):137-154, Nov 1992.
- [25] S. Wachter and H.-H. Nagel. Tracking of persons in monocular image sequences. *Proc. IEEE Nonrigid and Articulated Motion Workshop*, 1997.

## A. Orthographic 3D Reconstruction

In this appendix we present a factorization algorithm for the special case where the  $y$  image coordinates of a particular feature point are equal across multiple camera views. This case is a direct consequence of our use of Hu moments, which are invariant to translation, scaling, and rotation in the image plane. Since the SMA maps were trained with cameras with coinciding up-vectors, the  $v$  coordinates of the 2D estimates of a particular joint obtained from the inference are expected to be equal across multiple views.

The inputs to the algorithm are 2D joint locations estimates in the different virtual camera views obtained from

the SMA inference. The outputs are estimates of the camera's orientation and 3D joint estimates. The algorithm consists of the following steps:

**Step 0:** Normalize the data: (i) subtract the coccyx (tail bone) joint from all other joints. (ii) isotropically scale all but the first virtual camera view so that image scan lines match across all camera views.

**Step 1:** Construct the measurement matrix  $W$ :

$$W = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1N} \\ u_{21} & u_{22} & \dots & u_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ u_{C1} & u_{C2} & \dots & u_{CN} \end{bmatrix}, \quad (16)$$

where  $u_{ci}$  is the  $u$  image coordinate of feature point  $i$  in camera view  $c$ , and compute the singular value decomposition  $W = USV^T$ .

**Step 2:** The affine camera and structure reconstruction are:

$$P_c^a = \begin{bmatrix} U_{c1} & 0 & U_{c2} \\ 0 & 1 & 0 \end{bmatrix} \quad (17)$$

$$\begin{bmatrix} x_i^a & z_i^a \end{bmatrix} = S' * \begin{bmatrix} V_{i1} & V_{i2} \end{bmatrix}, \quad (18)$$

where  $S'$  is the  $2 \times 2$  major of  $S$ . Note that  $y_i^a$  can be set to the average of the  $v$  coordinates of feature point  $i$  across all camera views.

**Step 3:** Compute the rectifying homography by imposing the metric constraint, that is, find  $H$  such that for each camera  $c$

$$\mathbf{X}^e = H^{-1} \mathbf{X}^a \quad (19)$$

$$P_c^e = P_c^a H, \quad (20)$$

where  $\mathbf{X}^e$  is the Euclidean structure and  $P_c^e$  is the  $c$ -th Euclidean camera matrix:

$$P_c^e = \begin{bmatrix} a_c & 0 & b_c \\ 0 & 1 & 0 \end{bmatrix}. \quad (21)$$

Imposing the metric constraints

$$a_c^2 + b_c^2 = 1, \quad (22)$$

and the known form of the homography

$$H = \begin{bmatrix} h_1 & 0 & h_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (23)$$

and combining Eq. (19) - (23) for all  $C$  cameras yields a set of  $C$  linear equations in the unknowns  $h_1^2 + h_3^2$  and  $h_3$ :

$$[U_{c1}^2 \quad 2U_{c1}U_{c2}] * [h_1^2 + h_3^2 \quad h_3]^T = 1 - U_{c2}^2 \quad (24)$$

A solution for  $h_1$  and  $h_3$  can be readily obtained.

**Step 4:** Plug  $H$  in Eq. (20) and Eq. (19) to obtain Euclidean camera and structure.

**Step 5:** Align the first camera reference system with the world reference system by rotating its reference system into the identity matrix. Adjust the other cameras by using the same rotation.