Bayesian Articulated Tracking Using Single Frame Pose Sampling

Leonid Taycher and Trevor Darrell
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, 02139
{lodrion,trevor}ai.mit.edu

Abstract

We propose a novel probabilistic tracking framework for articulated bodies that incorporates direct estimation of the pose posterior distribution. We derive a single frame articulated pose sampler, and perform Bayesian tracking over time via Monte Carlo integration. In contrast to traditional particle filtering approaches, which propagate individual samples through time and are sensitive to the sample distribution, we generate samples directly from the current observation. Our method has the initialization benefits of single frame pose detection approaches, and stability benefits of sequential Monte Carlo methods. We have experimented with simple 2D head, hand, and contour edge observation likelihoods; our method is able to infer a distribution of full articulated pose from these simple features.

1. Introduction

Recovering human pose from image data is an important computer vision problem with applications in such areas as human-computer interaction, surveillance and content-based database retrieval.

Pose changes in sequential image data are commonly estimated using differential trackers [20, 1, 2, 3]. These trackers combine dynamics, prior pose information and the current frame data to estimate pose (or a pose distribution) for the current frame. This approach suffers from several common drawbacks, most critically error accumulation over time and the need for manual initialization. A complimentary approach is to track using the repeated application of a single image pose estimation technique at every frame [5, 17, 13, 18]. However, these methods do not use the pose information from previous frames and only estimate a single "best" pose that corresponds to the current observed image. Sequences of such estimates do not always correspond to correct dynamics due to the ambiguities that arise from projecting 3D bodies onto 2D images.

Our probabilistic tracking framework incorporates features of both approaches. For a single frame, the distribution of articulated pose parameters is estimated from static observations; with multiple frames, pose posteriors are propagated through time using a Bayesian technique. Our framework uses information from the current observation early in the inference process which improves the tracking stability when a strong dynamics model is not available.

Since parametric modeling of pose distributions is not feasible, we represent them with weighed sample sets. In our system, single frame pose parameter distributions are estimated using an importance

sampling technique [11]. We represent image likelihood functions using a generative model of body appearance (described in Section 3). Proposal distributions are automatically constructed from image measurements, kinematic constraints, and parameter priors (Section 4). Pose distribution samples for the current frame are evaluated with respect to a sampled representation of the prior pose distribution, producing a pose posterior conditioned on all observed data (Section 5). Since the pose is sampled at each frame independently, our system does not require initialization and is able to gracefully recover from tracking failures. Propagation over time ensures the temporal continuity of the pose estimate.

2. Prior Work

The problem of analyzing human pose from a single image has been addressed in several contexts. A 2D model in a dynamic programming framework was used for detecting humans or estimating human pose in natural scenes [17, 5]. A learning approach was taken by [18] to infer body pose from segmented silhouettes. Shape context matching was used in [13] to automatically locate joint positions. A tracking-by-continuous-detection framework that used geometric hashing was presented in [22]. Since these approaches generally produce a single estimate, using them naively in a tracking algorithm may produce errors when the pose cannot be uniquely determined from the image.

In recent years, particle filtering techniques have been widely used for articulated pose tracking. The posterior probability distribution is generated using dynamics and noise models to propagate samples of the prior over time; new weights are assigned based on the image likelihood model. This approach suffers from several drawbacks, such as drift due to noise or rapid motions, and the necessity to use a large number of particles to faithfully represent distributions in high dimensional spaces. In such spaces, sample impoverishment [9] may prevent particle filters from tracking multimodal distribution for long periods of time.

Several methods have been proposed for dealing with large particle set sizes, such as hybrid Monte Carlo filtering [3], annealed particle filters [4], and partitioned sampling [10]. ICONDENSATION [6] addressed the initialization and drift problems by combining regular CONDENSATION propagation with direct sampling from the re-initialization prior (which is commonly modeled with a simple parametric distribution). The major difference between our approach and CONDENSATION variants is that the samples of the posterior distribution from the previous frame are only used for generating the pose prior and not a proposal distribution. Thus the dependence on precise knowledge of body dynamics is reduced.

We use a generative appearance model defined by a Bayesian network similar in spirit to ones used in [8] for 3D articulated tracking and in [7] for modeling interactions of multiple independently moving objects in 2D. The major contribution of this work is incorporation of inverse kinematics-based constraints [12] directly into the likelihood computation stage of inference process, as opposed to using them during state prior propagation as done in [21].

3. Human Upper Body Model

We model the human upper body with the articulated model in Figure 1. The model configuration at time t is described by parameter vector (ϕ, Θ^t), where $\phi \in \mathfrak{R}^7$ contains time independent metric parameters (neck and upper and lower arm lengths, body width, depth and length, and head size, and $\Theta^t = (\theta_1^t, \ldots \theta_6^t) \in \mathfrak{R}^{16}$ contains pose parameters (three rotational degrees of freedom at neck and each

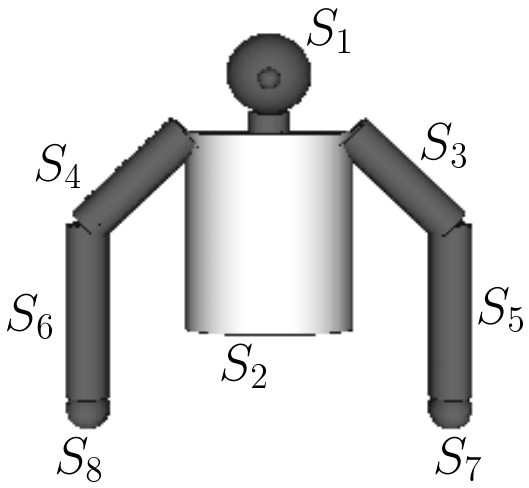


Figure 1: Articulated model of human upper body used in this work. The model consists of head (S_1), torso (S_2), and two arms with upper and lower arm segments (S_3-S_5 and S_4-S_6 respectively) and hands (S_7 and S_8). The model configuration includes 7 metric parameters: head radius, neck length, body width (distance between shoulder joints) length and depth, and upper and lower arm lengths. The pose is specified by 16 parameters: 11 internal rotational parameters (3 degrees of freedom at the neck, 3 degrees of freedom at each shoulder, and 1 at each elbow) and 5 global degrees of freedom (2 translational and 3 rotational).

shoulder, one at each elbow, and five global position parameters). Since we assume that the observed images are formed using orthographic projection, the global depth parameter is ignored.

The parameters of articulated joints $\{\theta_i^t\}$, metric parameters ϕ and segment appearances $\{A_i\}$ are modeled as independent. The pose of the ith segment, P_i, is deterministically computed from the pose of its parent in the articulated tree, denoted P_{p_i} and appropriate joint parameters θ_i. Segment appearance A_i and pose P_i are combined to produce a segment latent image L_i. Poses P_i are also used to compute binary support maps M_i for each segment (note that if the segment is not occluded, the support map depends only on the corresponding pose).

Due to the deterministic nature of the above steps, the following conditional pdfs used in the graphical model become delta functions:

$$\begin{aligned} p(P_i|P_{p_i},\theta_i,\phi) &= \delta(P_i - f_i^p(P_{p_i},\theta_i,\phi)) \\ p(L_i|P_i,A_i) &= \delta(L_i - f_i^l(P_i,A_i)) \\ p(M_i|\{P_i\}) &= \delta(M_i - f_i^m(\{P_i\})) \end{aligned} \tag{1}$$

where $\{P_i\}$ refers to $P_1, P_2, \ldots, P_8$, and f_i^p, f_i^l, f_i^m are functions that are used to compute the ith pose, latent image and support map, respectively.

The combined latent images (masked by their regions of support) corrupted by uncorrelated Gaussian noise are then observed as $\mathbf{I}$,

$$p(\mathbf{I}(x,y)|M_i,L_i) = N(\mathbf{I}(x,y); (\sum_i (M_i \cdot L_i)(x,y)), \sigma^2(x,y)) \tag{2}$$

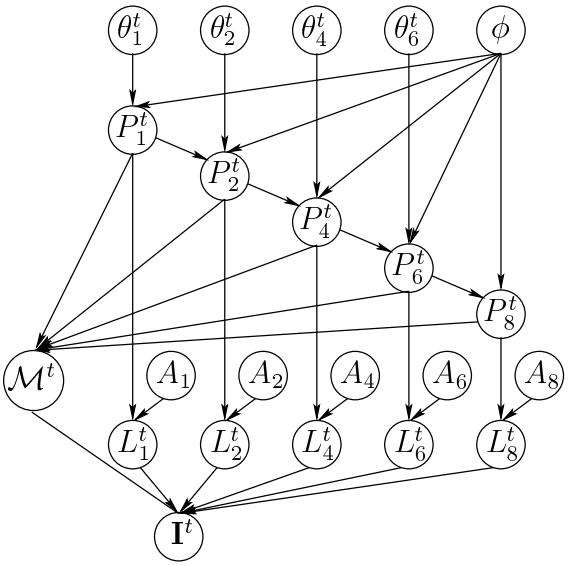


Figure 2: Generative model for an articulated body image (see Eq. 3). The subscripts correspond to segment number is Figure 1 (the nodes corresponding to S_3, S_5 and S_7 are symmetric to S_4, S_6 and S_8 and are not shown). The segment pose at time t, P_i^t is depends on the pose of a parent segment $P_{p_i}^t$, body lengths ϕ, and corresponding joint parameters θ_i^t. θ_1 contains the global position parameters. The latent image of a segment, L_i^t is obtained by transforming appearance A_i according to the pose P_i^t. The observed image $\mathbf{I}$ depends on the latent images masked by support maps $\mathcal{M}^t = (M_1^t, \cdots, M_8^t)$ that are, in turn, determined from all segment poses.

The joint probability of the described model may then be factored as

$$\begin{aligned} p(\phi,&\{\theta_i\},\{P_i\},\{A_i\},\{L_i\},\{M_i\},\mathbf{I}) = \\ &p(\phi)\prod_i p(\theta_i)\prod_i p(A_i)\prod_i p(P_i|P_{p_i},\theta_i,\phi) \\ &\prod_i p(L_i|P_i,A_i)\prod_i p(M_i|\{P_i\})p(\mathbf{I}|\{L_i\},\{M_i\}) \end{aligned} \tag{3}$$

This generative model is described by a graphical network in Figure 2.

In our model, the ith segment is "responsible" for the region of the observed image $\mathbf{I}$ that corresponds to its support map M_i. Let us define the ith observation region

$$\mathbf{I}_i = \mathbf{I}M_i = (\sum_i M_i L_i + \nu)M_i = L_i M_i + \nu M_i \tag{4}$$

Since support maps M_i are disjoint, the observation regions are independent, conditioned on $\{M_i\}$

and $\{L_i\}$. Furthermore, analyzing the conditional $p(\mathbf{I}_i|P_i, A_i, M_i)$, we find that

$$\begin{aligned} & p(\mathbf{I}_i|P_i, A_i, M_i) \\ &= \int_{L_i} p(\mathbf{I}_i|L_i, M_i)p(L_i|P_i, A_i) \\ &= \int_{L_i} p(L_i M_i + \nu M_i|L_i, M_i)\delta(L_i - f_i^l(P_i, A_i)) \\ &= \prod_{x,y \in M_i} N(\mathbf{I}_i(x,y); (f_i^l(P_i, A_i))(x,y), \sigma^2(x,y)) \end{aligned} \tag{5}$$

allowing us to use the simplified image generation model that induces the following joint pdf factorization

$$\begin{aligned} & p(\phi,\{\theta_i\}, \{P_i\}, \{A_i\}, \{M_i\}, \{\mathbf{I}_i\}) = \\ & \quad p(\phi) \prod_i p(\theta_i) \prod_i p(A_i) \prod_i p(P_i|P_{p_i}, \theta_i, \phi) \\ & \quad \prod_i p(M_i|\{P_i\}) \prod_i p(\mathbf{I}_i|\{P_i\}, \{A_i\}, \{M_i\}) \end{aligned} \tag{6}$$

This equation will be used for evaluating image likelihoods of poses generated using our single frame pose sampling framework

4. Sampling Articulated Pose

We with to infer a distribution of articulated pose parameters from a single frame. In this work, we assume that segment appearances $\{A_i\}$ and prior distributions of parameters $p(\theta_i)$ are known, and wish to describe the posterior distribution $p(\Theta|\{A_i\}, \mathbf{I})$. In the following discussion we assume that metric parameters of the model (ϕ_0) are also known and concentrate on sampling the posterior distribution of pose parameters $p(\Theta^t|I^t)$. We address estimation of ϕ in Section 6. Using Bayes' rule, independence assumptions, and Eq. 3 the pose posterior distribution may be expressed as

$$p(\Theta|\{A_i\}, \phi_0, \mathbf{I}) \sim p(\mathbf{I}|\Theta, \{A_i\}, \phi_0)p(\Theta) \tag{7}$$

The complexity of natural images makes it hard to specify this distribution analytically. While evaluating the posterior (up to a scaling factor) at any particular Θ_0 is relatively easy, sampling from it (necessary for tasks such as providing input to an articulated tracker) is hard. The alternative approach is to use Monte Carlo methods and represent $p(\Theta|\{A_i\}, \mathbf{I})$ as a set of samples with attached weights $\{\Theta_i, \pi_i\}$. One such method is importance sampling [11].

In the importance sampling framework, instead of sampling a target distribution $p(x)$, a *proposal* distribution $q(x)$ that approximates $p(x)$ is sampled, and then the weight of the sample x_k is set to $\pi_k = \frac{p(x_k)}{q(x_k)}$. A reasonable choice of a proposal distribution used in this technique should "concentrate" the samples in the areas of configuration space with high values of target distribution.

4.1. Proposal Distribution

Our approach to constructing a proposal distribution is based on the assumption that partial pose information for certain segments in the model may be extracted directly from the image. That is, it is possible

to efficiently sample from the conditional $p(\hat{P}_i|A_i, \mathbf{I})$, where $\hat{P}_i$ contains partial information about P_i. In our system such segments are head and hands (segments 1, 7, and 8). The appropriate models are discussed in Section 6.

We define our proposal distribution as

$$\begin{aligned} q(\Theta|A_1, A_7, A_8, \phi_0, \mathbf{I}) = \; & q(\Theta|\hat{P}_1, \hat{P}_7, \hat{P}_8, \phi_0) \\ & q(\hat{P}_1|A_1, \mathbf{I})q(\hat{P}_7|A_7, \mathbf{I})q(\hat{P}_8|A_8, \mathbf{I}) \\ = \; & q_{\text{head}}(\theta_1|\hat{P}_1, \phi_0)q_{\text{neck}}(\theta_2|\theta_1, \hat{P}_7, \hat{P}_8, \phi_0) \\ & q_{\text{left arm}}(\theta_3, \theta_5|\theta_1, \theta_2, \hat{P}_7, \phi_0) \\ & q_{\text{right arm}}(\theta_4, \theta_6|\theta_1, \theta_2, \hat{P}_8, \phi_0) \\ & p(\hat{P}_1|A_1, \mathbf{I})p(\hat{P}_7|A_7, \mathbf{I})p(\hat{P}_8|A_8, \mathbf{I}) \end{aligned} \tag{8}$$

We sample from q in five steps:

1. Obtain head and hands pose samples $\hat{P}_1^s$, $\hat{P}_7^s$, and $\hat{P}_8^s$ from the appropriate distributions.
2. Compute global parameters θ_1^s from $\hat{P}_1^s$.
3. Obtain neck joint configuration θ_2^s from q_{neck}
4. Obtain left arm configuration (θ_3^s and θ_5^s) by sampling from $q_{\text{left arm}}$
5. Obtain right arm configuration (θ_4^s and θ_6^s) by sampling from $q_{\text{right arm}}$

4.2. Kinematic Constraints

We would like to specify $q_{\text{neck}}(\cdot)$, $q_{\text{left arm}}(\cdot)$, and $q_{\text{right arm}}(\cdot)$ based on image information, joint parameter priors and kinematic constraints.

The samples of the distributions conditioned on the image, $\hat{P}_1^s = (\Omega, x_{p_1}, y_{p_1})^T$, $\hat{P}_7^s = (x_{p_7}, y_{p_7})^T$, and $\hat{P}_8^s = (x_{p_8}, y_{p_8})^T$ are the orientation and image position of the head and image plane coordinates hands (Section 6).

Without loss of generality we define the world coordinate system to have x and y axes parallel to image axes, and z axis passing through the origin of the image plane. Then the external parameters of the articulated model are simply $\theta_1 = (\Omega, x_{p_1}, y_{p_1})^T$, and $P_1^s = f_1^p(\theta_1, \phi_0)$.

Let us define a *feasible* configuration of shoulder pose $P_2 = f_2^p(\theta_2, f_1^p(\theta_1))$ and image plane hand locations $\hat{P}_7, \hat{P}_8$ to be one in which it is possible to reach each hand from a corresponding shoulder, that is, the image plane distance from the shoulder joint to the hand is less than the arm-length. Then, if we disallow all infeasible configurations, we can define q_{neck} as

$$\begin{aligned} q_{\text{neck}}(\theta_2|\theta_1, \hat{P}_7 = \hat{P}_7^s, \hat{P}_8 = \hat{P}_8^s, \phi_0) \\ \sim \begin{cases} p(\theta_2) & (f_2^p(\theta_2, f_1^p(\theta_1)), \hat{P}_7^s, \hat{P}_8^s) \text{ is feasible} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

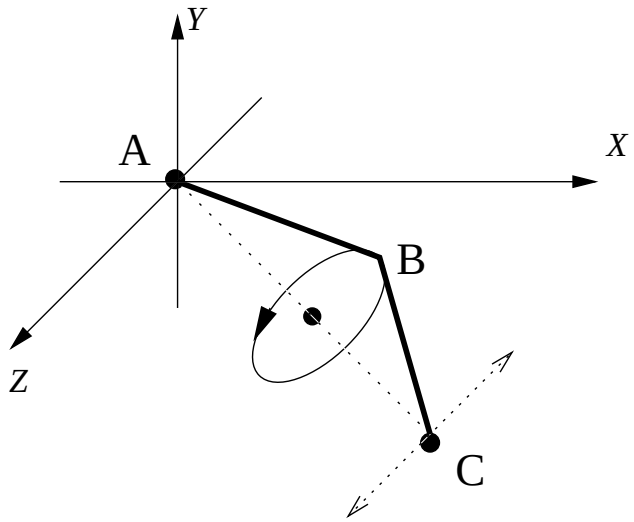


Figure 3: When the 3D position of the shoulder $\mathbf{A}$ and position of the hand $\mathbf{C}$ in the xy-plane are known, the arm has two degrees of freedom, depth of the hand and rotation of the elbow $\mathbf{B}$ about the shoulder-hand line.

Let us consider left arm as a two link assembly shown in Figure 3. The shoulders pose $P_2^s = f_2^p(P_1^s, \theta_2^s)$ uniquely determines the position of the left shoulder joint $\mathbf{A} = (x_A, y_A, z_A)^T$. The position of the left hand, $\mathbf{C}$ is known up to the translation along z axis

$$\mathbf{C} = (x_{P_7^s}, y_{P_7^s}, z_C)^T \tag{10}$$

$$z_A - r \leq z_C \leq z_A + r \tag{11}$$

$$r = \sqrt{(l_{\text{upper arm}} + l_{\text{lower arm}})^2 - (x_A - x_{P_7^s})^2 - (y_A - y_{P_7^s})^2}$$

where limits in Eq. 11 ensure that the distance between the shoulder joint and the hand is not greater than the total arm-length $(l_{\text{upper arm}} + l_{\text{lower arm}})$. The whole assembly may then be rotated about the line $\mathbf{AC}$ by $0 \leq \psi < 2\pi$. The configuration of the arm $\Theta_l = (\theta_3, \theta_5)^T$ is then *uniquely* determined by ψ, and z_C (i.e. $\Theta_l = g(z_C, \psi, P_1, \theta_2, \hat{P}_7)$). We model $q_{\text{left arm}}$ as

$$\begin{aligned} & q_{\text{left arm}}(\theta_3, \theta_5 | \theta_1, \theta_2, \hat{P}_7, \phi_0) \\ &= p(\theta_3, \theta_5 | z_C, \psi, \theta_2, \theta_2, \hat{P}_7, \phi_0) \\ & \quad p(z_C | P_1, \theta_2, \hat{P}_7, \phi_0) p(\psi | P_1, \theta_2, \hat{P}_7, \phi_0) \\ & \quad p(\theta_3) p(\theta_5) \\ &= \delta(\Theta_l - g(z_C, \psi, P_1, \theta_2, \hat{P}_7, \phi_0)) \\ & \quad p(z_C | P_1, \theta_2, \hat{P}_7, \phi_0) p(\psi | P_1, \theta_2, \hat{P}_7, \phi_0) \\ & \quad p(\theta_3) p(\theta_5) \end{aligned} \tag{12}$$

where

$$p(z_C | P_1, \theta_2, \hat{P}_7, \phi_0) = u(z_C; z_A - r, z_A + r) \tag{13}$$

$$p(\psi | P_1, \theta_2, \hat{P}_7, \phi_0) = u(\psi; 0, 2\pi) \tag{14}$$

The corresponding proposal distribution for the right arm, $q_{\text{right arm}}$, is defined in the same fashion. Once the sample $\Theta = (\theta_1, \ldots, \theta_6)$ is selected, we need to determine its weight $\pi = \frac{p(\Theta|\{A_i\},\phi_0,\mathbf{I})}{q(\Theta)}$. Note that

$$\begin{aligned} q_{\text{left arm}} \quad & (\Theta_3,\Theta 5|\theta_1,\theta_2,\hat{P}_7,\phi_0) \\ = & \int \delta(\Theta_l - g(z_C,\psi,P_1,\theta_2,\hat{P}_7,\phi_0)) \\ & p(z_C|P_1,\theta_2,\hat{P}_7,\phi_0)p(\phi_0|P_1,\theta_2,\hat{P}_7),\phi_0 \\ & p(\theta_3)p(\theta_5)dz_c d\psi \\ \sim & \begin{cases} p(\theta_3)p(\theta_5) & \text{if configuration is valid} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{15}$$

And thus, if we have obtained a sample $\Theta = (\theta_1, \ldots, \theta_6)$ from $q(\cdot)$, then

$$\begin{aligned} & q(\theta_1,\ldots,\theta_6|A_1,A_7,A_8,\phi_0,\mathbf{I}) \\ = \; & p(\hat{P}_1|A_1,\mathbf{I})p(\hat{P}_7|A_7,\mathbf{I})p(\hat{P}_8|A_8,\mathbf{I})\prod_{i=2}^{6} p(\theta_i) \end{aligned} \tag{16}$$

and the weight is given by

$$\begin{aligned} \pi = & \frac{p(\Theta|\{A_i\},\phi_0,\mathbf{I})}{q(\Theta\{A_i\},\phi_0,\mathbf{I})} \\ = & \frac{p(\theta_1)\prod_i p(\mathbf{I}_i|\{P_i\},\{A_i\},\{M_i\})}{p(\hat{P}_1|A_1,\mathbf{I})p(\hat{P}_7|A_7,\mathbf{I})p(\hat{P}_8|A_8,\mathbf{I})} \end{aligned} \tag{17}$$

By processing a frame $\mathbf{I}^t$ using algorithm described in this section, we obtain a sample set $\{(\Theta_i^t, \pi_i^t)\}$ representation of $p(\Theta^t|\mathbf{I}^t)$ that may then be used for tracking or estimating the Maximum Likelihood pose at the current timestep.

5. Pose Propagation Over Time

As has been discussed above, we would like to combine the pose estimate at the current frame with the previous observations, to produce a posterior distribution $p(\Theta^t|I^0 \ldots I^t)$. We make a Markovian assumption that all information about observations $I^0 \ldots I^t$ is preserved in distribution of Θ^{t-1}, that is $p(\Theta^t|\Theta^{t-1}, I^0 \ldots I^{t-1}) = p(\Theta^t|\Theta^{t-1})$. We then can express the full posterior as

$$\begin{aligned} p(\Theta^t|I^0 \ldots I^t) & \sim p(I^0 \ldots I^{t-1}|\Theta^t)p(I^t|\Theta^t)p(\Theta^t) \\ & \sim \frac{p(\Theta^t|I^t)}{p(\Theta^t)} \int p(\Theta^t|\Theta^{t-1})p(\Theta^{t-1}|I^0 \ldots I^{t-1})d\Theta^{t-1} \end{aligned} \tag{18}$$

In order to use the pose samples $\{\Theta_i^t\}$ used to represent $p(\Theta^t|I^t)$ to represent the full posterior, we need to compute the new weights λ_i^t. If we assume that the prior $p(\Theta^{t-1}|I^0 \ldots I^{t-1})$ is also represented

Algorithm 1 Sampling based articulated pose tracking

for all $t \geq 0$ **do**
  EXTRACT image features such as face position and flesh-colored blobs from the frame I^t.
  CONSTRUCT a proposal distribution $q^t(\Theta^t)$ from the extracted features and pose parameter priors.
  GENERATE N^t samples $\{(\Theta_i^t, \pi_i^t)|1 \leq i \leq N^t\}$ from the proposal distribution with
    corresponding weight computed as $\pi_i^t = p(\Theta_i^t|I^t)/q(\Theta_i^t)$.
  if the prior $p(\Theta^{t-1}|I^0 \dots I^{t-1})$ is available **then**
    GENERATE samples $\{(\Theta_i^t, \lambda_i^t)|1 \leq i \leq N^t\}$ from $p(\Theta^t|I^0 \dots I^t)$ by evaluating
      $\lambda_i^t = \pi_i^t \sum_{j=1}^{N^{t-1}} \lambda_j^{t-1} p(\Theta^t = \Theta_i^t|\Theta^{t-1} = \Theta_j^{t-1})/p(\Theta_t^i)$
    UPDATE $p(\phi)$ from $\{(\Theta_i^t, \lambda_i^t)|1 \leq i \leq N^t\}$
  else
    USE $\{(\Theta_i^t, \pi_i^t)\}$ as the estimate of $p(\Theta^t|I^0 \dots I^t)$
  end if
end for

with a set of weighted particles $\{(\Theta_j^{t-1}, \lambda_j^{t-1})\}$ obtained at the previous iteration of the algorithm, λ_i^t may be computed as

$$\lambda_i^t = \frac{\pi_i^t}{p(\Theta_t^i)} \sum_{j=1}^{N^{t-1}} \lambda_j^{t-1} p(\Theta^t = \Theta_i^t|\Theta^{t-1} = \Theta_j^{t-1}) \tag{19}$$

The complete tracking algorithm is presented in Algorithm 1.

6. Implementation

The description of our algorithm is completed by specification of the parameter priors $p(\theta_i)$, appearance A_i, and image formation models. We also need to address recovering metric model parameters ϕ.

We have obtained the joint angle limits from [14], and have represented shoulder and elbow angle prior probabilities as uniform between those limits. The neck angle prior was specified as a broad Gaussian centered on the origin.

For our method to be practical, the image formation models $p(\mathbf{I}_i|P_i, A_i, M_i)$ have to be efficient to evaluate, and, in the case of head and hands, lead to simple-to-sample-from posteriors $p(\hat{P}_i|\mathbf{I}, A_i)$. Many general techniques are possible. Here we use simple implementations flesh color and face patter detection.

Our general framework requires the estimate of the head pose $p(\hat{P}_1|\mathbf{I}, A_1)$. Ideally, we would use a face detector that is capable of detecting faces that have orientations other that frontal, while reporting size, location and orientation. For most of the experiments in this paper (other than Figure 5(b)) we assume that the person is in the upright position facing the camera, so we estimate $p(\hat{P}_1|\mathbf{I}, A_1)$ based on the output of a 2D frontal face detector (we use the method described in [23]). The detector output is a set of image squares that are reasonably well centered on the faces, and the distribution is represented as a mixture of Gaussians,

$$p(\hat{P}_1|\mathbf{I}, A_1) = \sum_f N(\hat{P}_1; c_f, \begin{pmatrix} 0.05r_f^2 & 0 \\ 0 & 0.05r_f^2 \end{pmatrix}) \tag{20}$$

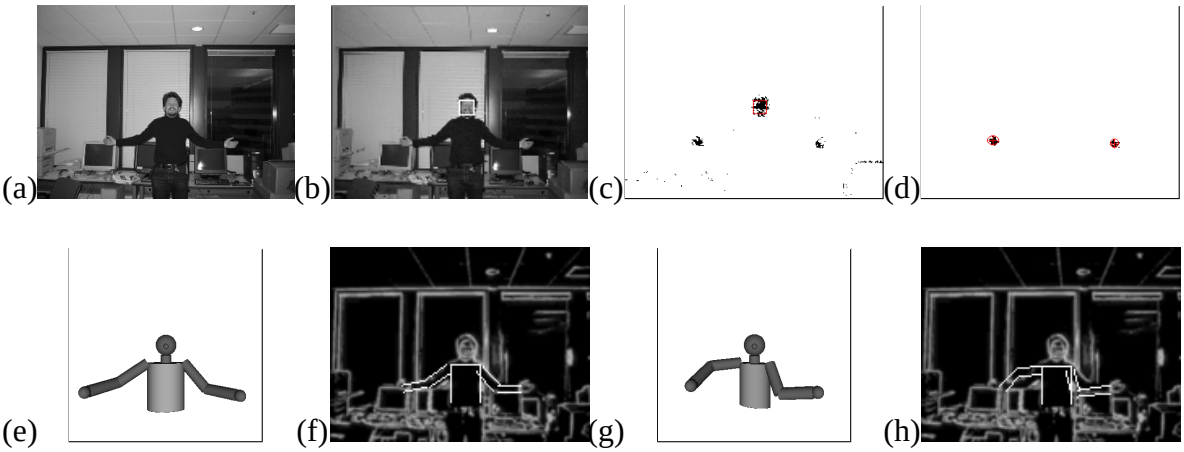


Figure 4: Stages of processing input image (a). The face rectangle (b) was located using a face detector [23], and the flesh color map (c) was computed by a detector initialized from the color distribution in the face rectangle . The result (d) of filtering the raw binary map (Section 6) was used to initialize the hand position distribution. Two sample poses with corresponding test edges overlaid on gradient image are shown in (e, f) and (g, h). The pose (e) was determined to be more likely that (g).

where c_f is a center of the detected square, and r_f is half of its width.

The face square size is also used to estimate the distribution of the metric parameter vector ϕ. We use anthropometric data from [14] combined with empirically estimated ratio of r_f to head radius to define means and standard deviations of Gaussian distributions from which elements of ϕ is drawn. Once the tracker has settled in, we replace the original metric prior $p(\phi)$ by the new prior $p_e(\phi)$ that is computed from the body sizes estimated the previous frames.

We model hands as flesh-colored blobs. The flesh color segmentation is performed on the input image using detector initialized from the middle region of the face rectangle (Figure 4(c)). Connected components are then computed from the resulting binary image. All components that either overlap the face rectangle, are larger than it in one of the dimensions, or have area smaller than 10% of the face rectangle area are filtered out. The hand pose posterior $p(\hat{P}_7|\mathbf{I}, A_7) = p(\hat{P}_8|\mathbf{I}, A_8)$ is then approximated as mixture-of-Gaussians where each constituent Gaussian distribution is initialized from one of the remaining components (Figure 4(d)).

We model elongated segments in the model (torso, lower and upper arms) as cylinders, and use the intensity gradient as image measurement. Along the contour of the segment's image plane projection (cf. Figure 4(f, h)), we expect the gradient to be perpendicular to the edge, and to have high magnitude [15, 16]. Let $\mathbf{G}_i^\alpha$ be the gradient direction image, E be the set of the points on the predicted edges under the support map, and α_0 be the predicted gradient direction. Then we define the image likelihood function as an average match along the predicted edges,

$$p(\mathbf{I}_i|P_i, A_i, M_i) \sim \frac{1}{|E|} \sum_{(x,y)\in E} e^{\cos 2(\mathbf{G}_i^\alpha(x,y)-\alpha_0)} \tag{21}$$

Sampling from Gaussian and uniform distributions is implemented using direct methods [11]. The distributions defined in Eqs. 9 and 12 are sampled by discretizing the parameter space, assigning each discrete sample s_i weight w_i proportional to the value of the appropriate pdf ($q_{\text{neck}}(s_i \dots)$ or $q_{\text{arm}}(s_i)$

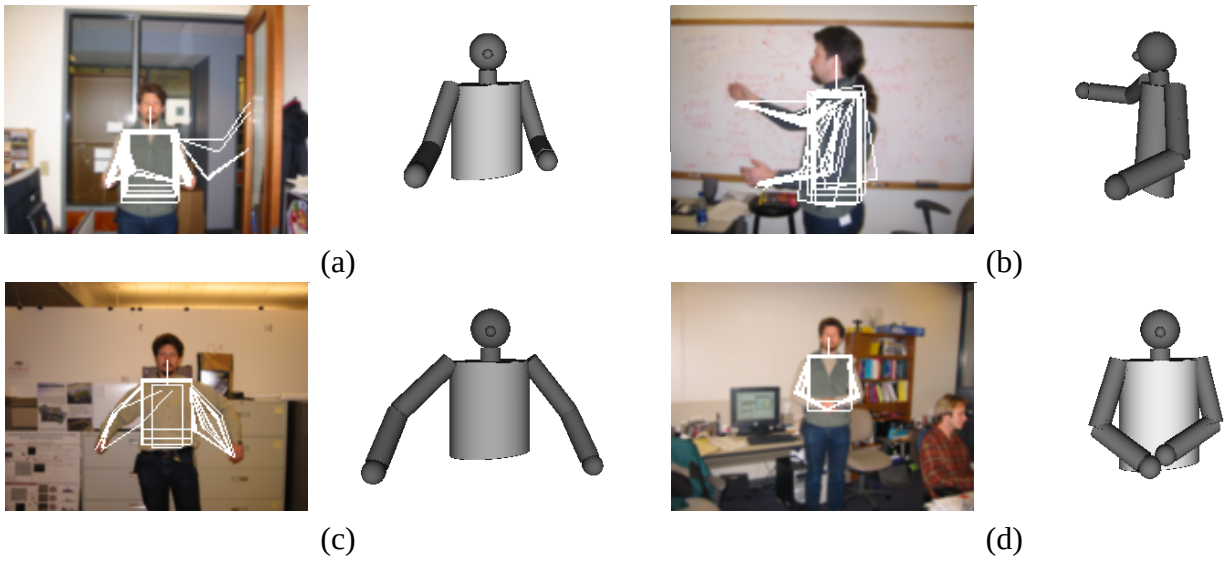


Figure 5: Sampling pose from a single image. For each example, twenty samples from the estimated posterior are shown overlaid on the image and the maximum likelihood pose is shown in 3D.

respectively), and then drawing sample from produced weighted sample set $\{(s_i, w_i)\}$. The state propagation probability is modeled using diffusion dynamics.

7. Results

We have applied our algorithm to a set of images of people in natural settings. Various stages of the algorithm are shown in Figure 4. The face rectangle detected in the input image (a) is shown in (b). Raw flesh color segmentation results and filtered image used to construct hand position distribution are (c) and (d). Panes (e) and (f), and (g) and (h) contain sample pose and corresponding edges overlaid on gradient magnitude image.

The results of applying our single frame pose detection algorithm to a set of four images is shown in Figure 5. For each of the examples, we present 20 random samples from the posterior pose distribution overlaid over the source image and the 3D reconstruction of the maximum likelihood particle. The head region and global transformation for the profile view (b) were manually initialized. Despite gross estimation errors in some samples, reporting a *range* of poses as opposed to a single result allows a higher level process to use additional information (such as motion or context) to select the most appropriate one.

An example of the algorithm's failure is shown in Figure 6. The image likelihood computation is confused by the strong background gradients, which results in incorrect pose estimation.

We have applied our tracking algorithm to the video sequence in Figure 7. The selected frames with poses sampled from estimated posterior are shown in top row. For comparison, the bottom row contains sample poses estimated using a simple CONDENSATION implementation (using diffusion dynamics). While our algorithm was able to successfully track through the whole sequence, strong drift and sample impoverishment have crippled CONDENSATION after the 50th frame.

Our system is currently implemented in in unoptimized C++. The total running time for a single frame while drawing 1000 samples (a number that has been empirically determined to be sufficient to represent pose distribution for these examples) requires, on average, three seconds.

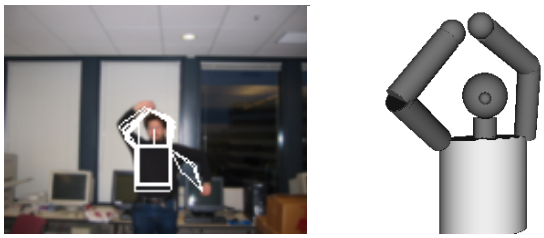
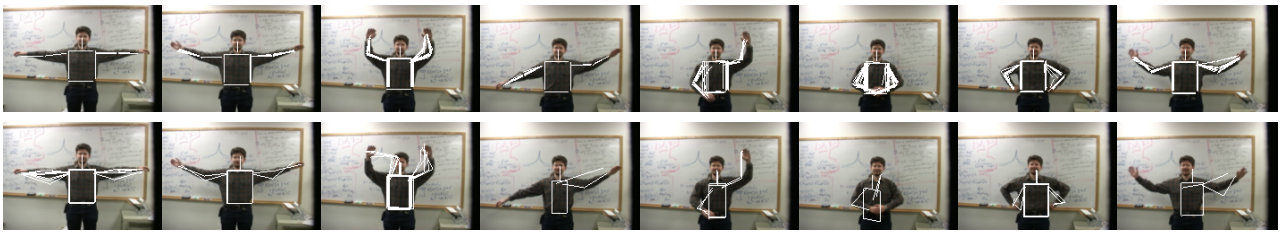


Figure 6: An example of the failure of pose estimation on a still frame. The algorithm was confused by the strong background image gradients, which resulted in assigning high probabilities to incorrect poses. Use of dynamic information in video mitigates such errors.



Frame 0 Frame 50 Frame 100 Frame 150 Frame 200 Frame 250 Frame 300 Frame 350

Figure 7: Applying our algorithm (top row) and CONDENSATION (bottom row) to a motion sequence. While our algorithm was able to track the body for the duration of the sequence, the CONDENSATION based tracker began drifting after the 50th frame.

8. Conclusions and Future Work

We have presented a technique for sampling human upper body pose posterior distribution from single images, and its application to tracking. In our approach, the kinematic constraints and image information are incorporated at early stages of inference process, which allows us to reduce the number of samples needed to approximate the high-dimensional articulated body distributions.

We use importance sampling with proposal distribution is constructed from prior probabilities of joint angles obtained from anthropometric data, inverse kinematics constraints, and from image face and hand locations detected by well-known methods. The observation likelihoods are represented using a novel Bayesian network description of a generative appearance model that also explicitly incorporates kinematic constraints. The distribution is propagated in time using Bayesian methods and Monte Carlo integration.

While our system behaves relatively well in moderately cluttered backgrounds, such backgrounds may confuse the simple segment appearance model that is used (e.g. Figure 6). Incorporating rich segment models of [19] or stereo input would alleviate this problem. Another enhancement would be incorporation of stronger dynamics model and more advanced sampling techniques such as hybrid Monte Carlo filters [3] applied both to samples estimated for the current image and propagated from the prior.

References

[1] Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. In *Proc. of CVPR*, 1998.

[2] Tat-Jen Cham and James M. Regh. A mutiple hypothesis approach to figure tracking. Technical report, Compaq Cambridge Research Laboratory, 1998.

[3] Kiam Choo and David J. Fleet. People tracking using hybrid monte carlo filtering. In *Proc. ICCV*, 2001.

[4] Jonathan Deutscher, Andrew Blake, and Reidm Ian. Articulated body motion capture by annealed particle filtering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.

[5] Pedro Felzenszwalb and Daniel Huttenlocher. Efficient matching of pictorial structures. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 66–73, 2000.

[6] Michael Isard and Andrew Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *ECCV (1)*, pages 893–908, 1998.

[7] Nebojsa Jojic and Brendan J. Frey. Learning flexible sprites in video layers. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[8] Nebojsa Jojic, Matthew Turk, and Thoman S. Huang. Tracking self-occluding articulated objects in dense disparity maps. In *Proc of International Conference on Computer Vision*, 1999.

[9] O. King and David A. Forsyth. How does CONDENSATION behave with a finite number of samples? In *ECCV (1)*, pages 695–709, 2000.

[10] John MacCormick and Michael Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *ECCV (2)*, pages 3–19, 2000.

[11] D.J.C MacKay. Introduction to monte carlo methods. In Micael I. Jordan, editor, *Learning in Graphical Models*, Adaptive Computation and Machine Learning, pages 175–204. MIT Press, 1998.

[12] Thomas B. Moeslund and Erik Granum. Multiple cues used in model-based human motion capture. In *FG'00*, 2000.

[13] Greg Mori and Jitendra Malik. Estimating human body configuration using shape context matching. In *European Conference on Computer Vision*, 2002.

[14] NASA. *Man-Systems Integration Standards Handbook*, 1995.

[15] James M. Rehg and Takeo Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. Fifth International Conference on Computer Vision*, pages 612–617, 1995.

[16] K. Rohr. Towards models-based recognition of human movements in image sequences. *CVGIP*, 59(1):94–115, Jan 1994.

[17] Remi Ronfard, Cordelia Schmid, and Bill Triggs. Learning to parse pictures of people. In *European Conference on Computer Vision*, Jun 2002. Copenhagen.

[18] Romer Rosales and Stan Sclaroff. Inferring body pose without tracking body parts. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 2000.

[19] Hedvig Sidenbladh. *Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequences*. PhD thesis, Royal Institute of Technology, Stockholm, 2001.

[20] Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Proc. European Conference on Computer Vision*, 2000.

[21] Christian Sminchiesescu and Bill Triggs. Kinematic jump processes for monocular 3d human tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[22] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *ECCV02*, page I: 629 ff., 2002.

[23] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision - to appear*, 2002.