Dynamic Human Pose Estimation using Markov chain Monte Carlo Approach

Mun Wai Lee and Ramakant Nevatia

Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, CA 90089-0273, USA
{munlee, nevatia}@usc.edu

Abstract

This paper addresses the problem of tracking human body pose in monocular video including automatic pose initialization and re-initialization after tracking failures caused by partial occlusion or unreliable observations. We proposed a method based on data-driven Markov chain Monte Carlo (DD-MCMC) that uses bottom-up techniques to generate state proposals for pose estimation and initialization. This method allows us to exploit different image cues and consolidate the inferences using a representation known as the proposal maps. We present experimental results with an indoor video sequence.

1 Introduction

Human body tracking is an important computer vision challenge in video understanding applications. This involves estimating the human poses in a given video sequence and it is useful for recognition of human gesture, analysis of human activities and understanding of human movement dynamics. This work focuses on human pose estimation with a monocular video that does not rely on markers on the human body. We use an articulated human model and the pose is represented by parameters that describe the body position, orientation and various joint angles.

1.1 Issues

The main issue in monocular human pose estimation is the lack of observables that can be used to infer the pose fully and reliably. The observations that are available contain noise, outliers, ambiguities, and do not directly provide depth information. There have been previous efforts that used multiple cameras, including range data from stereo [2], to obtain more reliable observable for pose estimation. However this framework is infeasible for many surveillance applications and for archived video analysis, where only monocular sequences are available.

The human pose is typically represented by more than 30 parameters. Therefore, an issue is how to search in this high dimension state space efficiently. In a video sequence, the problem can be formulated as one of dynamic state estimation. One may then use the previous state and a dynamic model to reduce the search space for the current state. Estimating the initial state is another key issue. In some sequences, there will be inter-occlusions of body parts during which the estimated states are unreliable. Therefore, there is a need to re-initialize (either a partial or the full state).

1.2 Related Work

There are various statistical techniques for tracking. For vision-based tracking of rigid objects, *Kalman filtering* technique is often used. For a nonlinear system, the *extended Kalman filter* (EKF) provides an alternative technique. EKF has previously been used for human pose tracking but it has severe limitations since it is a single mode tracker and breakdowns when the observation is ambiguous.

Particle filtering provides a powerful technique for estimating dynamic state with ambiguous observable, by approximating the state posterior distribution with a set of samples. But since human pose space is of high dimension, a particle filter is likely to degenerate when there are insufficient particles to characterize a complex state posterior distribution. Particle filtering has been used for pose tracking [3], but with poses in which the observation is relatively salient. Particle filtering may fail when, for example, there are inter-occlusions of body parts in past frames, unless there are other mechanisms to "relocate" the particles to particle-depleted regions of the state space [6].

In a 2D image, the depth of a joint is not directly observed; even if the image positions of two adjacent joints are known, there is a two-fold ambiguity in their relative depths. In [10], a mixture density propagation method is used to overcome depth ambiguities of articulated joints. This is later improved with a mixture smoother [11]. Recently, a *belief propagation* (BP) method has been used for pose estimation, in which the inference on each body component is propagated along a belief network [9]. This method requires the formulation of joint potential functions of cliques in the network, as well as conditional distributions of various observations. This is often difficult and approximating techniques have been introduced in a non-parametric BP framework. Nonetheless, there are limitations in BP, in regards to enforcing global constraints during evaluation.

To deal with the high dimensional state space and pose ambiguities, some previous efforts assume a low-dimensional dynamic space, (e.g. walking human)[1][13]. This scheme often results in good tracking in the studied movement, but less applicable for uncontrolled human motion. Recently, the use of bottom-up local parts detection is becoming a popular approach [8] to bootstrap the pose estimation.

1.3 Our Approach

We adapt a data-driven Markov chain Monte Carlo (DD-MCMC) approach [15] for dynamic pose estimation. In MCMC, pose estimation is formulated as a state estimation problem and the posterior is estimated by a set of samples generated as states in a Markov chain by a proposal function and accepted according to the Metropolis-Hastings algorithm.

This approach allows us to incorporate a generative human model that handles variations in human pose, body shape and simple clothing type. Given a state candidate, we can compute its likelihood measures by generating a synthesized human and comparing it to the input image.

The proposal function consists of different mechanisms to generate pose hypotheses. One mechanism is the forward propagation of previous frame pose estimates. Another mechanism is a bottom-up approach that exploits the observation in the current frame to infer the pose. The observation includes part-based detection of face, head-shoulder contours and limbs. We use a representation, known as *proposal map*, to consolidate the inference provided by the observation and to facilitate the data-driven proposal. It is not guaranteed that a proposal function derived from these observations will lead to the convergence to the true posterior. As these observations are used to reduce the search space, it may result in a proposal distribution with insufficient support for sampling the posterior. Using body part detectors that have high recall rates reduces the risk of this failure.

DD-MCMC approach and proposal maps have been previously used for human pose estimation in a single image in [5]; this work extends the method to estimate pose in a video sequence. Here, multiple Markov chains are used, one for each frame of the sequence and the estimation in adjacent frames are allow to influence the Markov transition based on a human dynamic model. Dynamic programming is used to extract the optimal trajectory of the human movement.

Our approach has two main advantages. First, it allows us to evaluate a pose candidate via synthesis using image domain cues such as region-based and color-based features, which are more easily obtained. In addition, evaluation is performed holistically after rendering the entire human form and this allows us to consider highly nonlinear constraints such as non-self-penetration. In contrast, evaluation approaches that examine each component independently often fail to consider such constraints.

Second, the DD-MCMC approach provides the flexibility to design different proposal mechanisms for generating state candidates and exploring the state space effectively. This includes using bottom-up techniques that exploit local component-level observations such as face detection [14] and limbs detection. This allows us to overcome the lack of observable in some frames and to recover smooth trajectories of the body poses.

This paper focuses on the case of only one person in the scene at a time. We believe that the framework can be extended to handle multiple persons and other objects (such as vehicles), including situations when they may inter-occlude; because with a top-down generative approach, hypotheses of multiple objects can be evaluated jointly. In this work, we have assumed an orthographic projection model, and that the person's face and hand skin regions are visible.

2 Estimation Framework

In this section, we describe the key components of our estimation framework including the human model, the observation, and the formulation of the prior distribution and likelihood function.

2.1 Human Model

The human model is an explicit representation of the human body structure. It defines the pose parameters which consist of the torso position, orientation and various body joint angles. Additional latent parameters are also included in order to synthesize the human appearance more suitably for pose evaluation. These latent parameters describe the shape of limbs and clothing type. Figure 1 illustrates the main components of the human model. More details are available in [5].

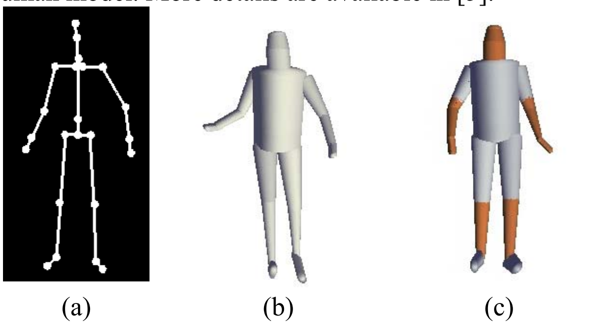

(a) (b) (c)

Figure 1: Human Model. (a) Kinematics structure, (b) shape, (c) cloth/skin visibility

2.2 State Estimation

Pose estimation is formulated as the problem of estimating the state of a system. Suppose we have a sequence with T frames, the state is represented by $\{X_1, X_2, ..., X_T\}$, where X_t represents the state at the t^{th} frame. The state includes the pose parameters (i.e. body position, orientation and joint angles), shape parameters and clothing parameters.

The shape and clothing parameters of a person are normally considered as static. However, since they are unknown and need to be estimated from observation, we consider these parameters as dynamic so that we can evaluate their likelihood measures by processing data within a local temporal neighborhood.

The observed images are denoted as $\{I_1, I_2, ..., I_T\}$. We assume that each observed image is conditionally dependent on the current state only. The relationships among the variables can be represented by a graphical model, as shown in Figure 2.

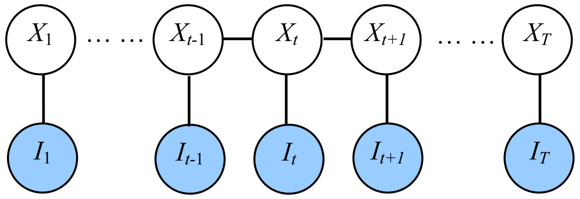


Figure 2. Graphical Model.

2.3 Prior Distribution

The prior distribution of the state, denoted by $p(X_1, X_2, ..., X_T)$, can be decomposed into a series of potential functions corresponding to the pairs of adjacent states in the graphical model.

$$p(X_1, X_2, ..., X_T) = \frac{1}{z} \prod_{t=1}^{T-1} \psi(X_t, X_{t+1}),$$

where Z is a normalization constant. Each potential function can be further decomposed into a prior distribution of a state, and a conditional distribution:

$$\psi(X_t, X_{t+1}) = p(X_t) p(X_{t+1} \mid X_t).$$

The prior distribution is learned from a training set of human poses in static image and sets of motion capture data. The conditional distribution is based on a zeroth-order dynamic model and is approximated by a normal distribution.

$$p(X_{t+1} \mid X_t) \approx N(X_{t+1} - X_t, \Sigma), \qquad (1)$$

where Σ is the covariance matrix of the dynamic model and is learned from motion capture data.

2.4 Observation and Likelihood Function

Observations are used in two complementary ways. First, the observations are used to evaluate pose candidates by synthesizing the human form and comparing it with the input image; this is described in this section. Second, some observations are used to design proposals for generating pose candidates during the Markov chain search. This is described in Section 3.

A state candidate X_t is evaluated by a likelihood function denoted by $p(I_t \mid X_t)$. Here, we assume that the image is conditionally dependent on the current state only.

We formulate the image likelihood function as consisting of four components, based on (*i*) region coherency, (*ii*) color dissimilarity with background, (*iii*) skin color and (*iv*) foreground matching. We describe them briefly in the following; the first two were also used in previous work and are described in more detail in [5].

Region coherency. Color-based segmentation is used to divide a given image into a set of regions. For a given state candidate, we predict the *human body region* in the image. Ideally, this human region will coincide with the union of a certain subset of the segmented regions. This region likelihood function measures the degree of similarity between the human body and the segmented regions.

Color dissimilarity with background. This likelihood measures the dissimilarity between the color distributions of the synthesized human region and the background region.

Skin color likelihood. The system state includes parameters describing the length of the sleeves. Therefore, given a state hypothesis, we can predict the positions of the visible skin regions and non-skin regions. This likelihood measures the likelihood of color in these regions based on their predicted types. Denoting c_i as the color value of the i^{th} pixel in the predicted human region, the likelihood is given by

$$L_{skin} = K_{skin} \left\{ \prod_{i=1}^{N} \left[P_{skin}(c_i) \right]^{l_i} \times \left[P_{non-skin}(c_i) \right]^{1-l_i} \right\}^{1/N},$$

where K_{skin} is a constant and N is the number of pixels in the predicted human region. $P_{skin}(c_i)$ and $P_{non-skin}(c_i)$ are the likelihoods of the pixel value for skin and non-skin region, derived from a histogram-based color model. l_i is a binary variable and has the value of 1 if the i^{th} pixel is the skin (based on the predicted pose), and has the value of 0 otherwise.

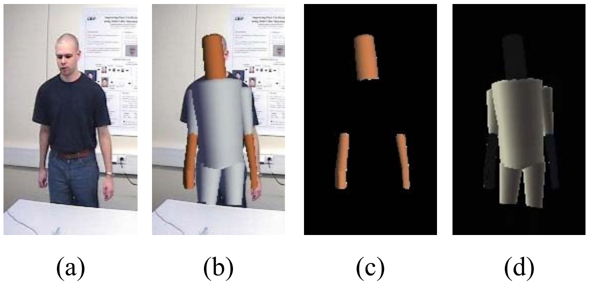


Figure 3. Skin Color Likelihood. (a) Input image, (b) rendered human, (c) predicted skin color regions, (d) predicted non-skin regions

Foreground matching. In some video sequences, we can extract the foreground region by background subtraction. This likelihood component measures the matching between the foreground and the synthesized human region based on overlapping ratio [16].

The combined likelihood measure is the product of the four likelihood components, assuming that they are independent.

3 Proposal Mechanisms

In this section, we describe the different proposal mechanisms used for the Markov chain transitions.

3.1 Proposal Function

The MCMC uses a proposal function to generate state candidates. There are techniques that generate samples that update the whole state sequence $\{X_1, X_2, ..., X_T\}$ (e.g. [7]), but such schemes have high computation complexity. Instead, we focus on the proposal function for a state X_t at a time.

The proposal function allows the Markov chain search to explore the state space efficiently. Due to the high-dimensional state space, the proposal function should be guided by evidence that provides inference of the state either fully or partially

We denote X_t' is the current Markov chain state, and X_t^* as a candidate for the new state. This candidate is generated by three types of evidence:

1. The estimation of previous state, X_{t-1} , can be propagated using a human dynamic model to generate candidates for the current state. This is often used in a sequential estimation framework and particle filtering is one specific technique. However, by itself, this is insufficient for dynamic pose estimation because inter-parts occlusion of the body would cause the state estimation to be unreliable during these frames.

2. The candidates can be generated from the current observation, I_t. This is an adaptation of a bottom-up data-driven approach [15] that has now been used for a number of computer vision tasks [12][16][17]. We shall discuss this in more details in Section 3.2.

3. Using a backward-propagation approach, the estimation of the next state, X_{t+1} , can also be used to generate candidates for the current state. We note that the backward-propagating component, $q(X_t^* \mid X_{t+1})$, is optional, depending on application requirement and constraint. This proposal mechanism is helpful in extracting a smooth pose trajectory and is applicable in a fixed-lag or batch mode of operation.

The proposal distribution is denoted by $q(X_t^* \mid X_{t-1}, I_t, X_{t+1}, X_t')$, where X_t' is the current Markov chain state. For simplicity, the distribution can be decomposed into its components:

$$q(X_t^* \mid X_{t-1}, I_t, X_{t+1}, X_t') = \alpha_1 q(X_t^* \mid X_{t-1}) + \alpha_2 q(X_t^* \mid I_t) + \alpha_3 q(X_t^* \mid X_{t+1}) + \alpha_4 q(X_t^* \mid X_t')$$

where $\alpha_1, \alpha_2, \alpha_3$, and α_4 are the mixing ratios for the different components. The last component, $q(X_t^* \mid X_t')$, represents a proposal distribution derived from the current Markov state. It is implemented to involve both the *random-walk sampler* [4], and the *flip kinematic jump* [10] which is specifically designed to explore the depth space for resolving the depth ambiguity.

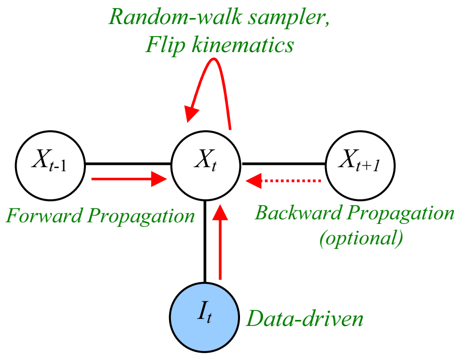


Figure 4. Types of Proposals.

3.2 Data-Driven Proposals

We use observations in the current frame to design bottom-up proposal for the state, $q(X_t^* \mid I_t)$. The observation include face detection, head-shoulder contour matching and skin blob extraction [5] (see Figure 5).

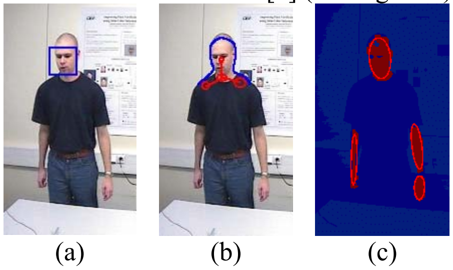


Figure 5. Image Observation. (a) Face detection, (b) head-shoulder contour matching, (c) skin blobs.

The evidence provided by these observations can be integrated and represented in *proposal maps.* There is one map for each body joint (Figure 6). In each map, the value at each pixel position represents the importance sampling probability of the corresponding joint's image position. The maps are used to generate pose candidate in a component-based Metropolis-Hastings approach. This technique is motivated by the data-driven MCMC framework [15][16]. In [5], it is shown how this framework can be adapted for estimating 3D kinematics parameters by constructing reversible jumps using the proposal maps and inverse kinematics computation.

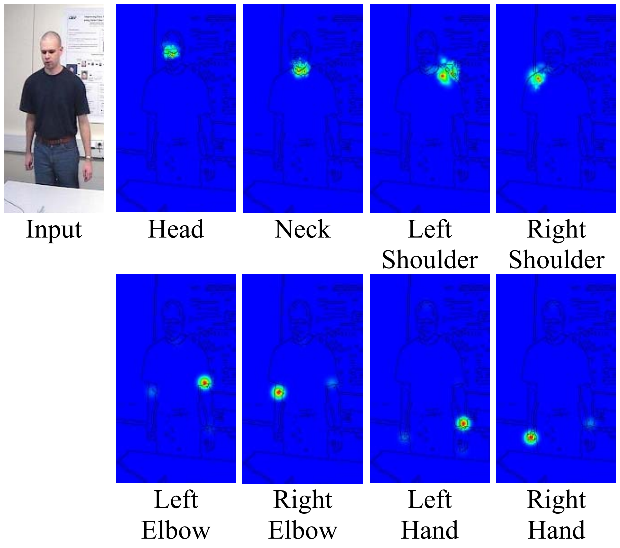


Figure 6. Proposal Maps, (shown in pseudo-colors).

3.3 Dynamic Proposals

Dynamic proposal mechanism involves generating a state candidate for the current frame, X_t^*, either from the estimates in the previous frame, X_{t-1}, or in the next frame, X_{t+1}. For discussion, we focus on the former.

The state estimation in the previous frame is represented by a set of state samples $\{X_{t-1}^1, X_{t-1}^2, ...\}$ generated by the Markov chain search. In practice, the number of state samples is typically about 500. We can approximate this by a mixture model approach with a more compact set of representative samples $\{X_{t-1}^{(1)}, X_{t-1}^{(2)}, ..., X_{t-1}^{(N)}\}$, where N is the number of mixture components. These components, obtained by clustering the samples, can be viewed as representing distinct modes in the posterior distribution, and they are weighted according to their cluster sizes.

To generate a state candidate for the current frame, a sample $X_{t-1}^{(*)}$ is selected from the set of mixture components in the previous frame $\{X_{t-1}^{(1)}, X_{t-1}^{(2)}, ..., X_{t-1}^{(N)}\}$ according to their normalized weights $\{w_{t-1}^{(1)}, w_{t-1}^{(2)}, ..., w_{t-1}^{(N)}\}$. The sample is then propagated using the dynamic model. Using a zeroth-order dynamic model, the state candidate is generated by sampling a normal distribution centered at $X_{t-1}^{(*)}$:

$$X_t^* \sim N(X_{t-1}^{(*)}, \Sigma),$$

where Σ is the covariance matrix of the dynamic model, as described in Equation (1) of Section 2.3.

The dynamic proposal distribution is given by:

$$q(X_t^* \mid X_{t-1}) = \sum_{1}^{N} w_{t-1}^{(i)} N(X_{t-1}^{(i)}, \Sigma).$$

4 Optimal Sequence of Pose

In Section 3, we have discussed how the estimation of the state for each frame is performed separately. This is equivalent to having multiple Markov chains, one for each frame.

In each frame, the set of generated Markov samples can be represented compactly using a mixture model as described in Section 3.3. The trajectory of the human body can be recovered by "traversing" along the sequence and selecting a smooth set of poses from these mixture components, this is implemented using dynamic programming.

For each frame, we denote θ_t as an index to one of the mixture components. We want to extract a set of indices $\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_T\}$ that represent the optimal sequence:

$$\{\hat{\theta}_1, ..., \hat{\theta}_T\} = \arg\max_{\{\theta_1, ..., \theta_T\}} \left\{ \prod_{t=1}^{T} p(I_t \mid X_t^{(\theta_t)}) \times \prod_{t=1}^{T-1} \psi(X_t^{(\theta_t)}, X_{t+1}^{(\theta_{t+1})}) \right\}$$

where $\psi(X_t^{(\theta_t)}, X_{t+1}^{(\theta_{t+1})})$ represents a potential function described in Section 2.3. It can be rewritten as:

$$\{\hat{\theta}_1,...,\hat{\theta}_T\} = \arg\max_{\{\theta_1,...,\theta_T\}} \left\{ \prod_{t=1}^{T} p(I_t \mid X_t^{(\theta_t)}) \times \prod_{t=1}^{T-1} p(X_t^{(\theta_t)}) p(X_{t+1}^{(\theta_{t+1})} \mid X_t^{(\theta_t)}) \right\} .$$

The dynamic programming equations are:

$$f_1(\theta'_1) = p(X_1^{(\theta'_1)}) p(I_1 \mid X_1^{(\theta'_1)}) ,$$

$$f_t(\theta'_t) = \max_{\hat{\theta}'_{t-1}} \left\{ p(I_t \mid X_t^{(\theta'_t)}) p(X_t^{(\theta'_t)}) p(X_t^{(\theta'_t)} \mid X_{t-1}^{(\theta'_{t-1})}) f_{t-1}(\theta'_{t-1}) \right\} ,$$

where $f_t(\theta'_t)$ is the joint posterior probability of the optimal sequence of poses from frame 1 to *t*, that terminates at the mixture component $X_t^{(\theta'_t)}$.

5 Experimental Results

In this section, we describe the experimental setup and the result.

5.1 Setup

A PETS-ICVS 2003 "smart meeting" video sequence [18] was used for evaluating the proposed technique. The video is annotated manually to aid in evaluation. An annotator located the image positions of the body joints. The depths of these joints, relative to the hip, were also estimated (Figure 7). The annotation data are used for evaluation only and not for training.

In this sequence, there are periods when the right arm is occluded. For these frames, the annotator is asked to estimate the positions of the occluded joints, based on his or her best judgment of the person's movement. The scene contains a table in the foreground that occludes the person's lower body. For this work, we manually extract the region of this table and consider it as an *a priori* known scene data. When the human model is rendered in the scene for evaluation, the region that is occluded by the table is ignored.

5.2 Dynamic Pose Estimation

The results of pose estimation are shown in Figure 9. The tracking started after the human fully entered the scene. The initialization is automatic. The shape of the human model was initialized arbitrarily and subsequently adapted to the input.

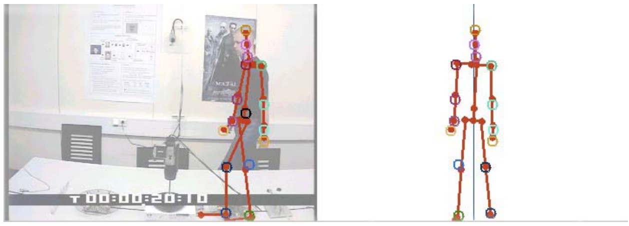
For each frame, about 500 Markov state samples were generated, requiring about 4 minutes per frame on average on a 2.8 GHz Intel P4 processor. We have not investigated the minimum number of samples necessary to obtain good estimation and instead generated a fairly large number of samples for each frame. We expect that the number of samples depends on the human movement and degree of ambiguity in the image, and that it varies significantly for different frames. For evaluation, we compare the estimated joint position with the annotated position in the image. We also compare the estimated depth (relative to the hip) with the manually judged depth. As we assume an orthographic projection, the depth is also expressed in pixel units.

In the t^{th} frame, we compute the 3D Euclidean distance error (in pixels) for the j^{th} joint, denoted by e_t^j . We then compute the RMS error for that frame, denoted by E_t , given as:

$$E_t = \left[\frac{1}{M} \sum_{j=1}^{M} (e_t^j)^2 \right]^{1/2} ,$$

where *M* is the number of joints used for evaluation. For this sequence, we evaluate based on 11 upper body joints, namely the hip, head, neck, shoulders, elbows, wrists, hands. These are used in all the frames, including the ones where part of the right arm is occluded.

The RMS error at each frame is plotted in Figure 8 (we call the testing of our proposed method Trial A) and the averaged RMS error for the whole sequence is given in Table 1. For comparison, we also conducted two other trials. In Trial B, the back-propagation proposal was not used, and in Trial C, neither the forward nor the backward propagation proposals were used. (In all the trials, dynamic programming was still used to extract the optimal sequences).



(a) Image view (b) Depth view

Figure 7. Pose Annotation. This is a snapshot of the pose annotation program. The colored circles indicate the joint positions located by the annotator, both on the image (a) and in the depth view (b). The depths are relative to the hip, which is always at the center of (b). To assist the user, a human stick figure is rendered to provide anthropometric reference. Here, the annotator had imagined where the occluded feet are, although the knees and feet were not used for the experimental evaluation.

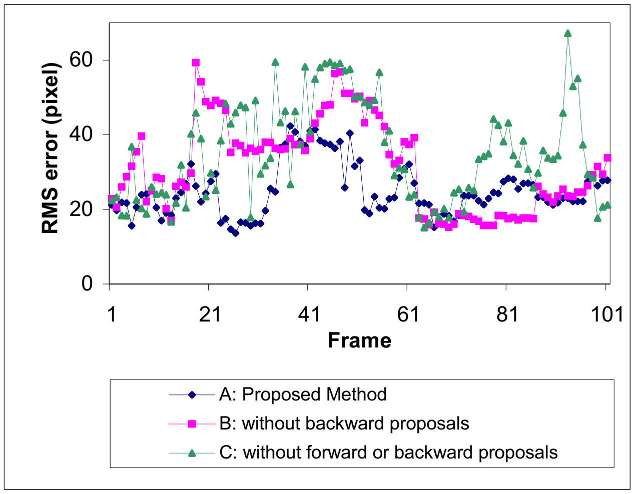


Figure 8. Pose Estimation Error

		Averaged RMS Error (pixels)
A	Proposed Method	24.99
B	without backward proposals	31.45
C	without forward nor backward proposals	35.31

Table 1. Average RMS Error for Sequence.

The results show that in some frames, single frame analysis with DD-MCMC is capable of obtaining good pose estimation, and these are the frames where error measures are comparable in all the three trials. In other frames, however, the errors are much higher for Trial B and C, when compared to Trial A. This indicates that the dynamic proposals (both forward and backward) are useful in providing more consistent pose estimation.

Between Frame 36 and Frame 49, the errors in Trial A are also high; this is one of the periods when the right arm is occluded. As the occlusion period is long, smoothing is unable to recover a good trajectory. Instead, the right hand is attracted to the skin region around the left elbow (see Figure 9(e,f)). We expect that some improvements in the likelihood function and the dynamic model will help overcome this problem. When the right arm reappears, our method is able to reestablish tracking, unlike what may be expected from many other tracking methods such as particle filtering, due to the data-driven proposal mechanism in our method.

6 Conclusion

We have proposed a data-driven MCMC method for dynamic pose estimation. This method allows us to incorporate different types of evidence for inferring body poses during the sequence. The evidence includes bottom-up image observation which enables the system to perform automatic pose initialization and re-initialization.

There are some limitations of the system. The MCMC is an iterative process and therefore more computationally expensive than methods such as belief propagation. Our current method for detecting limbs relies on skin color. In addition, the generative model currently handles only a single person.

We are currently exploring the use of shape-based features as alternatives for limb detection and improving the dynamic model so that the system can perform better during periods of partial occlusion.

Acknowledgment

This research was partially funded by the Advanced Research and Development Activity of the U.S. Government under contract: MDA-904-03-C-1786.

7 References

[1] K. Choo, and D.J. Fleet. People tracking with hybrid Monte Carlo, *ICCV 2001.*

[2] D. Demirdjian. Combining Geometric- and View-Based Approaches for Articulated Pose Estimation. *ECCV 2004*.

[3] J. Deutscher, A. Davison, I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture, *CVPR 2001.*

[4] W. Gilks, S. Richardson, D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[5] M. Lee, I. Cohen, "Proposal Maps driven MCMC for Estimating Human Body Pose in Static Images," CVPR 2004.

[6] M. Lee, I. Cohen, "Human Body Tracking with Auxiliary Measurements ", *IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003*.

[7] R. Neal, M. Beal, and S. Roweis. Inferring State Sequences for Non-Linear Systems with Embedded Hidden Markov Models. *NIPS 2003*.

[8] T. J. Roberts, S. J. McKenna, I. W. Ricketts: Human Pose Estimation Using Learnt Probabilistic Region Similarities and Partial Configurations. *ECCV 2004*.

[9] L. Sigal, S. Bhatia, S. Roth, M. J. Black, M. Isard, Tracking Loose-limbed People, *CVPR 2004*.

[10] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular Human Tracking, *CVPR 2003*.

[11] C. Sminchisescu and A. Jepson. Variational Mixture Smoothing for Non-Linear Dynamical Systems, *CVPR 2004*.

[12] Z.W. Tu and S.C. Zhu, Image Segmentation by Data-Driven Markov Chain Monte Carlo, *PAMI 24(5)*, pp. 657-672, 2002.

[13] R. Urtasun and P. Fua, 3D Human Body Tracking Using Deterministic Temporal Motion Models, *ECCV 2004*.

[14] P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features, *CVPR 2001.*

[15] S. Zhu, R. Zhang, Z. Tu. Integrating bottom-up/top-down for object recognition by data driven Markov chain Monte Carlo, *CVPR 2000*, vol.1, pp.738 –745.

[16] T. Zhao, R. Nevatia. Bayesian Human Segmentation in Crowded Situations, *CVPR 2003.*

[17] T. Zhao and R. Nevatia, Tracking Multiple Humans in Crowded Environment, *CVPR 2004.*

[18] http://www.cvg.cs.rdg.ac.uk/PETS-ICVS/

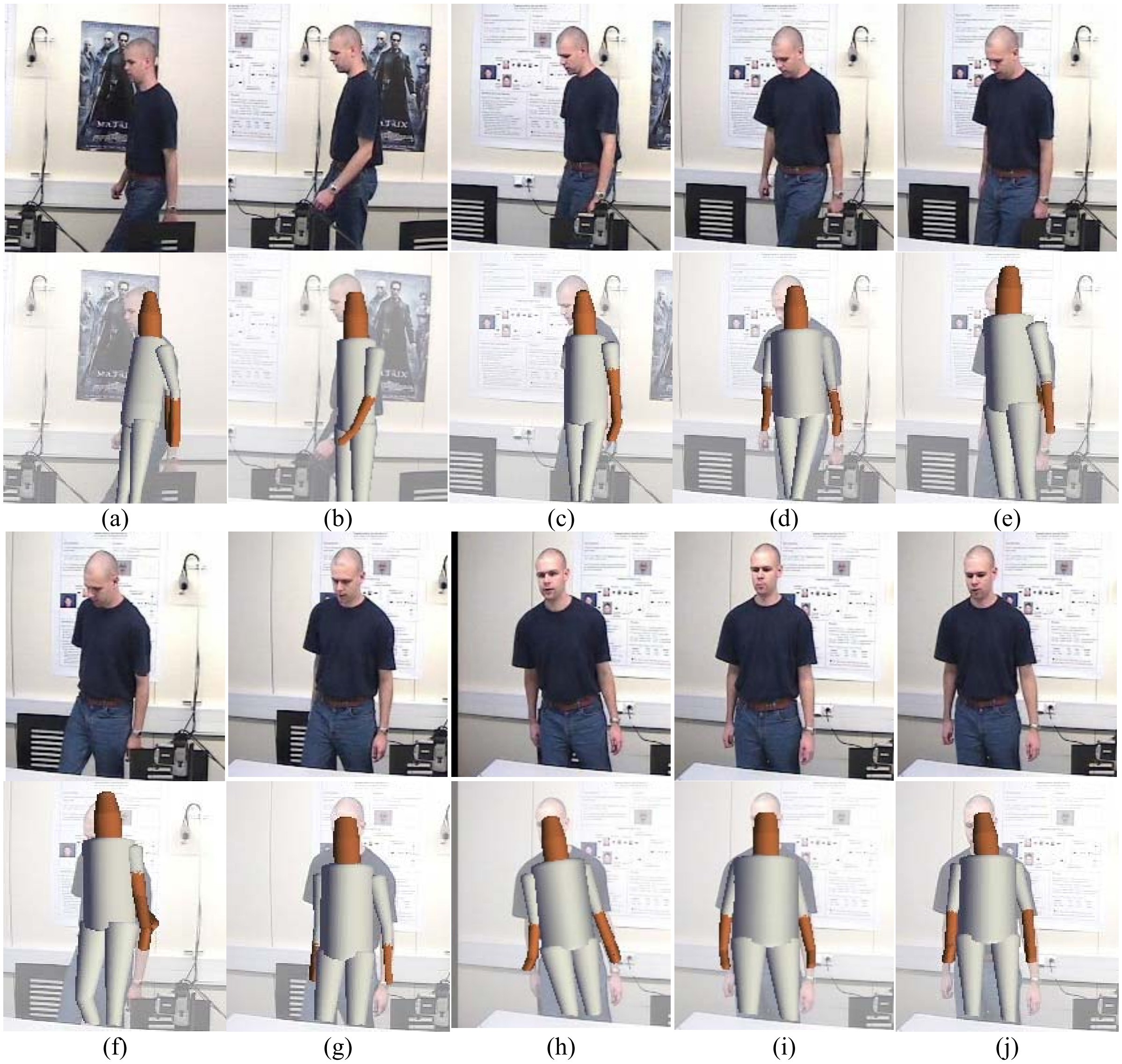


Figure 9. Dynamic pose estimation with PETS video with body turning movement. The person's right hand was occluded in (b) and (c), but its track was reestablished in (d) when it reappeared. There is also another period in (e) and (f) when the right hand was again occluded and the estimations were erroneous as the right hand was attracted to the skin regions around the left hand. Track was regained in (g).