Proposal Maps driven MCMC for Estimating Human Body Pose in Static Images

Mun Wai Lee and Isaac Cohen
Institute for Robotics and Intelligent Systems
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-0273, USA
{munlee, icohen}@usc.edu

Abstract

This paper addresses the problem of estimating human body pose in static images. This problem is challenging due to the high dimensional state space of body poses, the presence of pose ambiguity, and the need to segment the human body in an image. We use an image generative approach by modeling the human kinematics, the shape and the clothing probabilistically. These models are used for deriving a good likelihood measure to evaluate samples in the solution space. We adopt a data-driven MCMC framework for searching the solution space efficiently. Our observation data include the face, head-shoulders contour, skin color blobs, and ridges; and they provide evidences on the positions of the head, shoulders and limbs. To translate these inferences into pose hypotheses, we introduce the use of 'proposal maps', which is an efficient way of consolidating the evidence and generating 3D pose candidates during the MCMC search. As experimental results show, the proposed technique estimates the human 3D pose accurately on various test images.

1. Introduction

Estimating human body pose is important for automatic recognition of human activities in image understanding applications. For static images, the major difficulties are the high dimensionality of the solution space, pose ambiguity, and body segmentation.

The human body has about 31 parameters and pose estimation involves searching in a high dimensional and multi-modal solution space. In addition, there is an inherent non-observability of some of the degrees of freedom, causing "forwards/backwards flipping ambiguities" [10] in the depths of body joints. Ambiguity is also caused by noisy or spurious image features.

The segmentation of the human body is required because the human boundary in an image is dependent on the body pose, and this boundary affects the feature extraction needed to estimate the body pose. This calls for a method that simultaneously solves the dual-problem of segmentation and pose estimation.

We propose to address this problem by building an image generative model and using the Markov chain Monte Carlo (MCMC) framework [2] to search the 31-D solution space. Our human model represents the kinematics structure, shape and clothing of the human body. Given a pose candidate, a human image can be synthesized and compared with the real image. This model-based approach is appealing as it seeks to "*explain away the data"* from the image generation standpoint [14]; and it solves the segmentation problem simultaneously. The set of samples drawn by MCMC weakly converges to a stationary distribution equivalent to the posterior distribution. However, with only the use of random-walk sampler algorithm, the MCMC framework is inefficient.

The data-driven MCMC framework [14] allows us to design complementary *jump* proposal functions, derived from image observations, to explore the solution space more efficiently. Each jump dynamic has a much larger scope and allows the transitions between non-neighboring regions of high densities. Useful image observations are obtained from appearance-based face detection, matching head and shoulders contours, skin color blobs detection, and limbs detection. However, these observations only provide 2D inferences on local body parts but not the 3D pose. Also, these observations have localization errors, contain false alarms, and may not be independent. A mechanism is needed to properly translate these observations into proposal distributions for the 3D pose, while addressing the above issues.

In this paper, we describe the use of *proposal maps* to consolidate the inferences provided by the collective set of image observations. Each proposal map represents the proposal distribution of the image position of a body joint and it is used to generate proposals of 3D pose during MCMC.

We focused on middle resolution images, where the body height is about 150 pixels. We make no restrictive assumptions about the background, the human shape, and clothing except for not wearing any headwear or gloves.

The paper is organized as follows: Section 2 discusses related work; Section 3 describes the MCMC framework and its extension to pose estimation; the generative model and image observation are presented in Sections 4 and 5, respectively; and Section 6 shows experimental results.

2. Related Work

Pose estimation in video sequences has been addressed in many previous works, using either multiple

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)
1063-6919/04 $20.00 © 2004 **IEEE**

or a single camera [1][9]. Many of these works used particle filter to track the body poses, by relying on a good initialization, temporal smoothness, and sometimes a low dimensional dynamic model [1].

For static images, some works have been reported for recognizing prototypical body poses using shape context descriptors [5], mapping of features into body configurations [7], and parameter-sensitive hashing [8]. These works rely on either a clean background or a presegmented human region and not suitable for automatic pose estimation.

There are reported works on detecting body parts in images. In [6][3], the authors model the appearance and the 2D geometric configuration of body parts. These methods focus on real-time detection of people and do not estimate the 3D body pose. Recovering 3D pose was studied in [11], but the proposed method assumes that the image positions of body joints are known and therefore simplifies the problem.

3. Estimation Framework

3.1. Data-Driven MCMC

In this section, we describe the MCMC framework [2][14] and its adaptation for pose estimation. Denoting $\boldsymbol{x}$ as the vector of model parameters and $\boldsymbol{Y}$ as the image observations, pose estimation is formulated as a Bayesian inference for estimating the posterior distribution:

$$p(\boldsymbol{x}\,|\,\boldsymbol{Y}) \propto p(\boldsymbol{Y}\,|\,\boldsymbol{x})p(\boldsymbol{x})\,. \tag{1}$$

The desired output is dependent on the application. A simple solution is the maximum *a posteriori* estimate (MAP) given by:

$$\boldsymbol{x}_{MAP} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}\,|\,\boldsymbol{Y})\,. \tag{2}$$

Since the posterior distribution is multi-modal, it is often desirable to extract multiple solutions of $\boldsymbol{x}$. A simple approach consists in approximating the posterior distribution as a mixture of Gaussians:

$$p(\boldsymbol{x}\,|\,\boldsymbol{Y}) \approx \sum_i w_i N(\boldsymbol{\mu}_i, \Sigma)\,. \tag{3}$$

MCMC is a suitable methodology for finding these solutions by drawing samples from the posterior distribution using a Markov chain based on the Metropolis-Hastings algorithm [2]. At the t-1$^{\text{th}}$ iteration, a candidate $\boldsymbol{x}'$ is sampled from a proposal distribution $q(.\,|\,\boldsymbol{x}_{t-1})$ and accepted as the new state $\boldsymbol{x}_t$ with a probability $a(\boldsymbol{x}_{t-1} \to \boldsymbol{x}')$ where

$$a(\boldsymbol{x}_{t-1} \to \boldsymbol{x}') = \min\left\{1, \frac{p(\boldsymbol{x}'\,|\,\boldsymbol{Y})q(\boldsymbol{x}_{t-1}\,|\,\boldsymbol{x}')}{p(\boldsymbol{x}_{t-1}\,|\,\boldsymbol{Y})q(\boldsymbol{x}'\,|\,\boldsymbol{x}_{t-1})}\right\}\,. \tag{4}$$

A simple proposal distribution is provided by the *random-walk sampler* [2].

In the data-driven MCMC paradigm, image observations are used to generate additional proposals to enhance the convergence of the MCMC [14]. As the mapping of 2D image features to 3D poses is non-trivial, we design a proposal mechanism based on the missing data $\boldsymbol{u}$ that represents the image positions of body joints (head, elbows, etc.). The observation $\boldsymbol{Y}$ is conditionally dependent on $\boldsymbol{u}$ only, i.e. $p(\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{x}) = p(\boldsymbol{Y}|\boldsymbol{u})$. In addition, $\boldsymbol{u}$ can be computed from $\boldsymbol{x}$ by a deterministic function $\boldsymbol{u}=f(\boldsymbol{x})$ which is a many-to-one mapping. We can decompose $\boldsymbol{u}$ into its components $\boldsymbol{u}=\{\boldsymbol{u}_i\}$, where $\boldsymbol{u}_i$ is the image position of the i^{th} body joint, and use local observations, $\boldsymbol{Y}_i \subset \boldsymbol{Y}$, to generate the proposal distribution for this joint, denoted by $q_i(\boldsymbol{u}_i\,|\,\boldsymbol{Y}_i, \boldsymbol{x}_{t-1})$. The function $f(.)$ can be decomposed into its components $f(.)=\{f_i(.)\}$ such that $\boldsymbol{u}_i = f_i(\boldsymbol{x})$. (See Figure 1 for graphical models of these variables.)

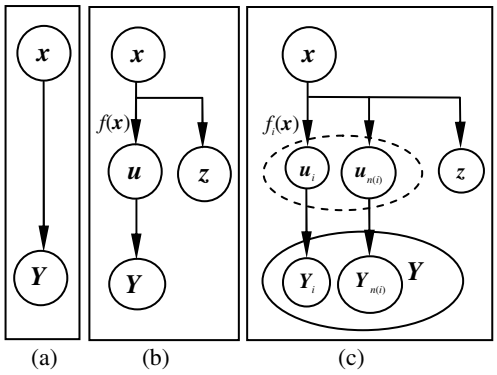


Figure 1: *Graphical Models: (a) shows the basic model between model parameters* $\boldsymbol{x}$ *and image observation* $\boldsymbol{Y}$*, (b) introduces the missing data* $\boldsymbol{u}$*, representing image positions of body joints, and* $\boldsymbol{z}$ *representing the depth, (c) shows relationship between the image position of the* i^{th} *body part* $\boldsymbol{u}_i$ *and the local observation* $\boldsymbol{Y}_i$*, which is used to generate proposals. In this figure,* $\boldsymbol{u}_{n(i)}$ *represents image positions of other body joints except the* i^{th}*, and* $\boldsymbol{Y}_{n(i)}$ *represents the corresponding local observations for these joints.*

Using a component-based proposal approach, the i^{th} component proposal distribution for $\boldsymbol{x}'$ becomes:

$$q_i(\boldsymbol{x}'\,|\,\boldsymbol{x}_{t-1}, \boldsymbol{Y}_i) = \int q_i(\boldsymbol{x}'\,|\,\boldsymbol{u}_i, \boldsymbol{x}_{t-1}, \boldsymbol{Y}_i) q_i(\boldsymbol{u}_i\,|\,\boldsymbol{x}_{t-1}, \boldsymbol{Y}_i) d\boldsymbol{u}_i\,. \tag{5}$$

The sampling of the proposal distribution is simplified in two ways. First, we construct the proposal for $\boldsymbol{u}_i$ so that it is independent of the previous sample $\boldsymbol{x}_{t-1}$:

$$q_i(\boldsymbol{u}_i\,|\,\boldsymbol{x}_{t-1}, \boldsymbol{Y}_i) = q_i(\boldsymbol{u}_i\,|\,\boldsymbol{Y}_i)\,. \tag{6}$$

Second, we construct the proposal $q_i(\boldsymbol{x}'\,|\,\boldsymbol{u}_i, \boldsymbol{x}_{t-1}, \boldsymbol{Y}_i)$ as a deterministic function. Given the previous state $\boldsymbol{x}_{t-1}$ and a sample $\boldsymbol{u}_i'$, a candidate $\boldsymbol{x}'$ is computed easily using direct inverse kinematics (IK) so that the image position of the i^{th} body joint is shifted to $\boldsymbol{u}_i'$, while all the other body joints are left unchanged:

$$f_j(\boldsymbol{x}') = \begin{cases} f_j(\boldsymbol{x}_{t-1}) & \text{when } j \neq i \\ \boldsymbol{u}_i' & \text{when } j = i \end{cases} \quad (7)$$

When multiple solutions exist, due to depth ambiguity, we choose the solution $\boldsymbol{x}'$ with the smallest change in depth. Due to space limitation, we will not discuss the inverse kinematics in details. Instead, readers may refer to [10] for a discussion. Denoting the IK computation as a function $g(.)$, we have:

$$\boldsymbol{x}' = g(\boldsymbol{x}_{t-1}, \boldsymbol{u}_i) . \quad (8)$$

The proposal distribution then becomes:

$$q_i(\boldsymbol{x}' \mid \boldsymbol{x}_{t-1}, \boldsymbol{Y}_i) = \int \delta(\boldsymbol{x}' - g(\boldsymbol{x}_{t-1}, \boldsymbol{u}_i)) q_i(\boldsymbol{u}_i \mid \boldsymbol{Y}_i) d\boldsymbol{u}_i , \quad (9)$$

where $\delta(.)$ is the Dirac delta function. We can draw a sample for $\boldsymbol{x}'$, by first drawing a sample $\boldsymbol{u}_i'$ from $q_i(\boldsymbol{u}_i \mid \boldsymbol{Y}_i)$, and computing $\boldsymbol{x}'$ using Equation (8). At each Markov chain iteration, this step is repeated for different body joint, in a partitioning approach.

Sometimes there is no valid IK solution due to kinematics and no self-penetrating constraints. (For example, the proposed hand position might be too far from the elbow.) In this case, the function $g(.)$ outputs an invalid state, denoted by $\boldsymbol{x}_{null}$, which has the property $p(\boldsymbol{x}_{null})=0$. The proposal driven by $\boldsymbol{u}_i'$ is then rejected according to Equation (4).

3.2. Proposal Maps

This section discusses the proposal mechanism for drawing a sample $\boldsymbol{u}_i \sim q_i(\boldsymbol{u}_i \mid \boldsymbol{Y}_i)$, where $\boldsymbol{Y}_i$ represents the observation of the i^{th} body joint. The observation usually includes false alarms and generates multiple weighted hypotheses on the image position of the i^{th} joint. We express $\boldsymbol{Y}_i$ as the set of hypotheses:

$$\boldsymbol{Y}_i = \{w_{i,k}, \boldsymbol{Y}_{i,k}; k = 1,..., n_i\} , \quad (10)$$

where $\boldsymbol{Y}_{i,k}$ represents each hypothesis, $w_{i,k}$ its confidence, and n_i is the number of hypotheses. The inference of each hypothesis is approximated by a Gaussian distribution with mean $\mu_{i,k}$ and covariance matrix $\Sigma_{i,k}$ corresponding to the measurement uncertainty. The proposal distribution for the joint image position derived from each hypothesis is given by:

$$q(\boldsymbol{u}_i \mid \boldsymbol{Y}_{i,k}) \propto N(\boldsymbol{u}_i, \boldsymbol{\mu}_{i,k}, \Sigma_{i,k}) , \quad (11)$$

and the contributions of all the hypotheses are combined as follows:

$$q(\boldsymbol{u}_i \mid \boldsymbol{Y}_i) \propto \max_k \{w_{i,k} \times q(\boldsymbol{u}_i \mid \boldsymbol{Y}_{i,k})\} . \quad (12)$$

The hypotheses are, in general, not independent. For example, matching the head-shoulders contour to image edges usually results in multiple local minima. In addition, the hypotheses are generated using different types of image cues and this leads to redundancy. We therefore use the *max* function in Equation (12) instead of the *summation* to avoid exaggerated dominant peaks.

We present in this section a method for improving sampling efficiency. The proposal distribution is approximated by a grid space representation called the *proposal map*, with samples corresponding to every pixel position. The proposal distribution is thus correctly bounded within the image. As this distribution is unchanged during the MCMC process, it is computed beforehand. Ignoring the quantization noise (which is small compared to the measurement errors), this proposal is reversible: for any valid proposal jump, there is another valid reverse jump because the function $g(.)$ in Equation (8) has a one-to-one mapping.

In Figure 2 we show the grey level representation of the proposal maps for various body joints. They are generated from image observations that we will describe in Section 5.

3.3. Other Proposal Mechanisms

The Markov chain dynamic consists of three types of proposals: (*i*) data-driven proposal, which was described earlier, (*ii*) random-walk sampler, and (*iii*) flip kinematics jump. The last two are briefly described here.

Random-walk Sampler. This process serves as a local optimizer and the corresponding proposal distribution is given by:

$$q(\boldsymbol{x}' \mid \boldsymbol{x}_{t-1}) \propto N(\boldsymbol{x}' - \boldsymbol{x}_{t-1}, 0, \Sigma_{diffusion}) . \quad (13)$$

Flip Kinematic Jump. This dynamic involves flipping a body part (i.e. head, hand, lower arm, entire arm, lower leg, or entire leg) along the depth direction, around its pivotal joint [10]. Flip dynamic is balanced so that forward and backward flips have the same proposal probability.

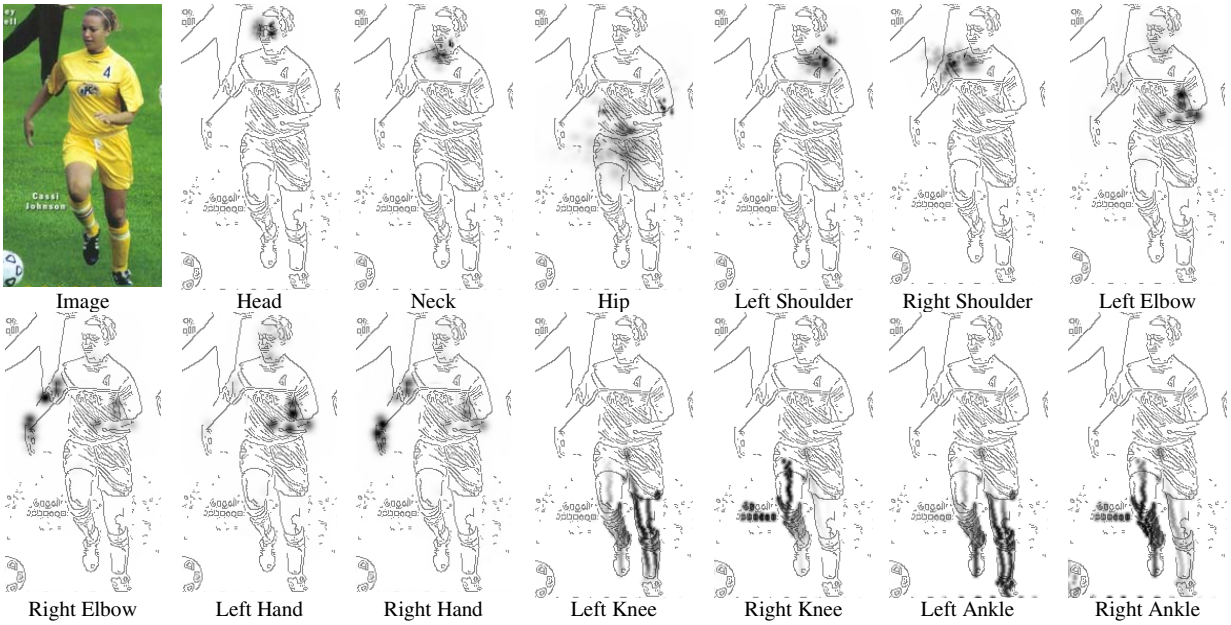


Figure 2: *Grey level representation of proposal maps for various body joints (overlaid on edge image for clarity).*

4. Generative Model

4.1. Human Model

The human model is an explicit representation of the human body structure. It defines the pose parameters as well as the parameters for shape and clothing.

Human Kinematics Model. This model represents the articulated structure of the human body and has 31 degrees of freedom. The pose is described by a 6D vector $\boldsymbol{g}$ representing global position, scale, and orientation, and a 25D vector $\boldsymbol{j}$ representing the joint angles. We assumed an orthographic projection. The prior distributions of these parameters, denoted by $p(\boldsymbol{g})$ and $p(\boldsymbol{j})$, are learned from the training data. For simplicity, these distributions are approximated as Gaussians and the joint angles of non-neighboring body locations are assumed to be independent.

Probabilistic Shape Model. Each human body part is represented by a truncated 3D cone and the shape of the human is represented by a vector $\boldsymbol{s}$ which has 23 parameters describing the relative lengths and widths of these 3D cones. PCA is used to reduce the shape space to 6 dimensions. The prior distribution $p(\boldsymbol{s})$ is assumed to be Gaussian.

Clothing Model. This model describes the type of clothing the person is wearing and allows the prediction of whether skin is exposed so that skin blob features are correctly interpreted. As there are many clothing types, the modeling requires a trade-off between generality and simplicity. The clothing model has 3 parameters, $\boldsymbol{c} = [c_1\ c_2\ c_3]^T$, representing the sleeve length, the hem length, and the socks length. For computation efficiency, these parameters are quantized into coarse discrete levels (5 levels for c_1, and 10 levels for c_2 and c_3). The prior distribution, $P(\boldsymbol{c})$, is learned from the training data.

4.2. Prior Distribution

The parameters from the various components of the human model are combined into a complete state vector $\boldsymbol{x}$, now consisting of four subsets:

$$\boldsymbol{x} = \{\, \boldsymbol{g}, \boldsymbol{j}, \boldsymbol{s}, \boldsymbol{c} \,\} \,. \tag{14}$$

For simplicity, we assume that the subsets of parameters are independent and the prior distribution, denoted by $p(\boldsymbol{x})$, is given by:

$$p(\boldsymbol{x}) \approx p(\boldsymbol{g})\, p(\boldsymbol{j})\, p(\boldsymbol{s})\, P(\boldsymbol{c}) \,. \tag{15}$$

This prior distribution is combined with the image likelihood function to form the posterior density function, which is used for evaluating samples and computing the acceptance probability for the Markov chain, as given by Equation (4).

The following sub-section describes the image likelihood function.

4.3. Image Likelihood Function

The image likelihood function $p(\boldsymbol{Y}|\boldsymbol{x})$ consists of two components, based on region and color respectively. This approach is motivated by the work in [15] where similar likelihood measure is used for segmenting multiple persons in static image.

Region Likelihood. Color-based segmentation is used to divide a given image into a set of regions denoted by

$\{R_i;\ i = 1, ..., N_{region}\}$, where N_{region} is the number of regions. For a given state candidate $\boldsymbol{x}$, we predict the *human body region* in the image, denoted by H_x. Ideally, this human region will coincide with the union of a certain subset of the segmented regions. In other words, each region R_i should either belong to the human region H_x or to the *background* (non-human) region, denoted by $\overline{H}_{\boldsymbol{x}}$. This region likelihood function measures the degree of similarity between the human body and the segmented regions. For each segmented region R_i, we count the number of pixels in R_i, that belong to H_x, and that belong to $\overline{H}_{\boldsymbol{x}}$:

$$\begin{aligned} N_{i,human} &= \text{count pixels } (u,v) \in \{ R_i \cap H_{x_t} \} \\ N_{i,background} &= \text{count pixels } (u,v) \in \{ R_i \cap \overline{H}_{\boldsymbol{x}} \} \end{aligned} \quad (16)$$

We define a binary label, l_i, for each region, so that

$$l_i = \begin{cases} 1 & \text{if } N_{i,human} \geq N_{i,background} \\ 0 & \text{otherwise} \end{cases} . \quad (17)$$

We count the number of *incoherent* pixels, denoted by $N_{incoherent}$, given by:

$$N_{incoherent} = \sum_{i=1}^{N_{region}} \left(N_{i,backgruond}\right)^{l_i} \left(N_{i,human}\right)^{(1-l_i)} . \quad (18)$$

The region-based likelihood function is defined by:

$$L_{region} = \exp(-\lambda_{region} N_{incoherent}) , \quad (19)$$

where λ_{region} (=0.2) is a constant determined empirically with training data using a Poisson model.

Color Likelihood. The likelihood measures the dissimilarity between the color distributions of the human region H_x and the background region $\overline{H}_{\boldsymbol{x}}$. Given the predicted region H_x, and $\overline{H}_{\boldsymbol{x}}$, we obtain the color distribution of human region $\boldsymbol{d}$, and background region $\boldsymbol{b}$. They are represented by normalized histograms with $N_{histogram}$ bins. The color likelihood is defined by:

$$L_{color} = \exp(-\lambda_{color} B_{\boldsymbol{d},\boldsymbol{b}}^2) , \quad (20)$$

where λ_{color} (=30) is a constant determined empirically and $B_{\boldsymbol{d},\boldsymbol{b}}$ is the Bhattachayya coefficient measuring the similarity of two color distributions given by:

$$B_{\boldsymbol{d},\boldsymbol{b}} = \sum_{i=1}^{N_{histogram}} \sqrt{d_i b_i} . \quad (21)$$

The combined likelihood measure is given by:

$$p(\boldsymbol{Y} \mid \boldsymbol{x}) = L_{region} \times L_{color} . \quad (22)$$

5. Image Observations

Image observations are used to compute the proposal maps described in Section 3. These are local observations used to infer positions of various body joints, and they are weighted according to their saliency and joint probabilities. These observations are extracted in 4 stages: (*i*) face detection; (*ii*) head-shoulders contour matching; (*iii*) skin blobs detection; and (*iv*) ridges detection.

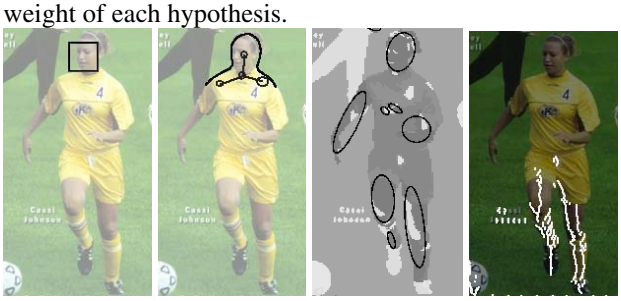
5.1. Face Detection

The AdaBoost cascade classification technique [13] is used for detecting faces in the image. Each detected face provides a hypothesis of the head position. As the face constitutes the most reliable observation, the detected face is used to initiate the extraction of other image observations.

5.2. Head-Shoulders Contour Matching

An active shape model approach is used to detect the head and shoulders contour using a deformable shape model. The face observation is used to define a search region within which multiple candidates for the head-shoulders contour are detected using a gradient descent search that aligns the shape model to image edges .

Each detected contour provides hypotheses on the positions of head, neck and shoulders (Figure 3.b), using a joint probabilistic model of these variables. The edge matching error is used to adjust the confidence weight of each hypothesis.



(a) (b) (c) (d)

Figure 3: *Image observations: (a) box indicates the detected face; (b) outline indicates one of the detected head-shoulders contour, ellipses indicate the corresponding hypotheses for head, left and right shoulders, and the ellipse size represents measurement uncertainty; (c) a grey-level map of skin probability with extracted skin ellipses; and (d) white pixels indicate ridges for the lower body.*

5.3. Elliptical Skin Blob Detection

Skin color features provide important cues about the positions of the arms and sometimes the legs. Skin blobs are detected in four sub-stages: (*i*) the image is divided into regions using a color-based image segmentation; (*ii*) for each segmented region, the probability of skin is evaluated using a histogram-based skin color model; (*iii*) ellipses are fitted to the boundary of these regions to form skin ellipse candidates; and finally (*iv*) adjacent regions with high skin probabilities are merged to form larger regions and extract larger ellipses (see Figure 3.c). The extracted skin ellipses are used for inferring the positions of limbs. Further details are given in [4].

5.4. Ridge Observations

In addition to skin color blobs, another type of observations useful for the segmentation of the limbs is based on ridges. If most of the limbs are clothed, especially the lower body, skin color blobs are less useful and ridge observations are more relied upon. The centers of the ridges provide hypotheses on the medial axis points of the limbs and therefore provide inference on the position of the legs. This approach is motivated by the work in [3] where limbs are detected as rectangular segments.

In the following, we discuss two aspects of the observations: (*i*) the detection of ridges, and (*ii*) the computation of confidence weights for these observations.

Detection of Ridges. We extract the medial axis points of image regions derived from color-based segmentation. Since the lower body limbs are usually not horizontal, the medial axis points are easily extracted by scanning each horizontal line in the image to find the centers of each region along the line. To overcome errors due to imperfect color segmentation, we first use an over-segmented image to find the first set of medial axis points. We then merge neighboring regions with similar color and extract additional medial axis points on the new regions. This method extracts many medial axis points efficiently.

Confidence weight. Each extracted point is weighted by a confidence measure based on the following criterions: (*i*) the likelihood of the point being on the medial axis of the leg, (*ii*) the likelihood of the region (to which the point belongs) being a subset of the leg, and (*iii*) the likelihood of the width of the region.

In order to compute these likelihoods, we need first a joint probability model, learned from training data, of the positions of the legs and the position of the torso. Estimates of the torso position are provided by head-shoulder contour matching.

The confidence measures are used to prune out some of the medial axis points with low confidence (Figure 3.d). We use the remaining points and the corresponding weights to generate the proposal maps for the knees and ankles, as described in Section 3.

Figure 2 shows examples of proposal maps for lower body joints. These distributions are generally more diffused, as the observations are less reliable. Because each observation could be associated to either side of the limbs, the maps for the left and right legs are similar to some extent. Nonetheless, the proposal maps do capture different regions of high densities to indicate plausible positions of legs and allow for *proposal jumps* to explore these regions during Markov chain iterations.

6. Experiments

Database and Ground Truth. We used a set of images representing various human activities on which we have generated ground truth by manually locating joint positions and estimating their relative depths. Among this set, we chose a subset for training (primarily for learning the prior distributions of model parameters), and the rest for testing the proposed method. This second subset is used for the experiments described in this section.

The experiments were conducted *without* any manual pre-processing such as background removal, scaling and centering of the person, or model initialization. At the start of the MCMC search, the human model was initialized in a standard upright pose in the center of the image.

Pose Estimation. Figure 4 shows the obtained 3D pose estimation on various images from the test set after 1000 Markov chain iterations. The estimated human model and its pose (the *MAP* solution) are projected onto the image and a 3D rendering from a sideward view is also shown to illustrate the depth estimation. Some small errors are observed and discussed in the figure caption; these are mostly due to the lack of image observable or features. The overall 3D pose estimation is good on this challenging set of images.

The estimated joint positions were compared with the ground truth data, and a RMS error was computed using all body joints. Since the depth has a higher uncertainty, we have computed two separate RMS errors: one for the 2D position and the other for the depth. We computed the average of these errors over all test images (*average RMS error*); the result (based on 20 images) is given in Table 1. As the posterior distribution is multi-modal, the *MAP* solution may be insufficient. As an alternative measure, we approximated the posterior distribution as a mixture model by clustering the samples using *k-mean* algorithm (we used k =20). (An alternative technique is the greedy "K-adventurers" algorithm [12] which updates the mixture model after each iteration.).

Figure 4: **Pose Estimation**. *1st Row: Original images, 2nd row: estimated poses, 3rd row: side view. Due to space limitation, the images were cropped for display.* ***Errors*** *include:* ***(1)*** *the person's left lower leg in image A, as it is mostly hidden,* ***(2)*** *the left arm in B, where the elbow is highly bent,* ***(3)*** *the left foot in C which is dark and similar to background,* ***(4)*** *the feet in E is wrongly estimated to be tip-toed. In addition, there are errors in depth estimates such as the right elbow in B, the tilting of the torso in B, the left feet in C, and the right arm in D.*

The clusters were *ranked* by their sample sizes. Using the estimated cluster means, *average RMS errors* were computed based on *Rank 5, 10, 15* criterions. (For example, *Rank 5* result was obtained by finding the lowest error among the five highest ranked cluster means.) These results, presented in Table 1, show that the mixture model captures better estimates of the pose, especially the depth estimates. Good pose estimates are usually found within the 5 highest ranked components.

		Average RMS Error (pixel)	
		(image position)	(depth)
MAP Solution		14.92	21.45
Mixture Model Solution	*Rank 5*	12.44	15.04
	Rank 10	11.94	14.86
	Rank 15	11.89	14.83

Table 1. *Average RMS errors* in image position and depth, using *MAP* solutions and *mixture model* solution.

To examine the spread of the RMS errors among test images, Figure 5 shows a histogram of these errors using *Rank 5* results.

Convergence Analysis. Figure 6 shows the RMS errors with respect to the MCMC iterations. The error for the 2D image position decreases rapidly from the start of the MCMC process; this is due largely to the observation-driven proposal dynamics. For the depth estimate, the kinematics flip dynamic is helpful for finding good depth estimates, but it requires a longer time for exploration. In the current implementation, 1000 iterations were considered and it took, on average, 8 minutes.

7. Conclusion

We have presented a data-driven MCMC framework for estimating 3D human pose in static images. Image observations of different cues provide inferences on the image positions of body joints. We introduce the use of *proposal map* as an effective mechanism to consolidate these inferences and generate 3D pose candidates for MCMC. As the results show, the technique is effective on a wide variety of images.

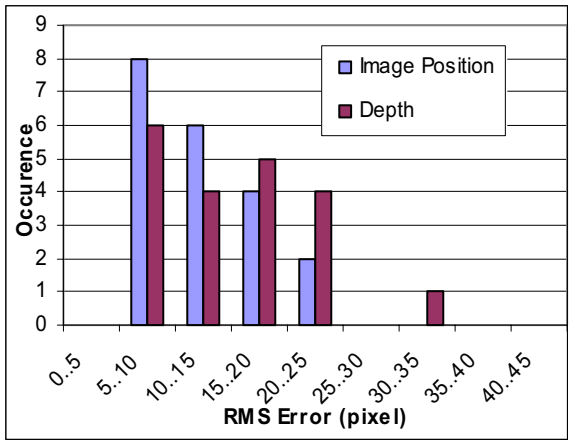


Figure 5: *Histogram of Average RMS Error (using Rank 5 result).*

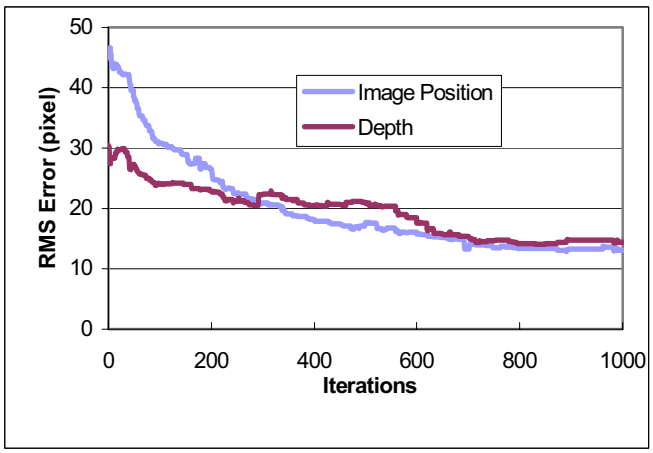


Figure 6: *Convergence Analysis (using Rank 5 result). Note that the relative depth estimate has no global offset and therefore has smaller error at the start of iteration compared to image position.*

The system currently has two main limitations. Firstly, the technique requires a good face detection algorithm. The face detection method used is reliable only for frontal faces. Secondly, the computational cost is still quite high, even with the use of data-driven proposal. As future work, we are exploring better techniques to detect non-frontal faces. In addition, we are designing techniques based on gradient-based diffusion and Gibbs sampling to improve the efficiency of the MCMC algorithm.

Acknowledgment

This research was partially funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, under Cooperative Agreement No. EEC-9529152, and by the Advanced Research and Development Activity of the U.S. Government under contract: MDA-908-00-C-0036.

References

[1] K. Choo, D.J. Fleet. People tracking with hybrid Monte Carlo, *ICCV 2001,* pp. 321-328.

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

[3] S. Ioffe and D.A. Forsyth. Probabilistic methods for finding people, *IJCV 43(1),* pp.45-68, June 2001.

[4] M. Lee, I. Cohen. Human Upper Body Pose Estimation in Static Images. *ECCV 2004.*

[5] G. Mori, J. Malik. Estimating Human Body Configurations using Shape Context Matching. *ECCV 2002*, pp 666-680.

[6] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. *ECCV 2002*, vol. 4, pp. 700-714.

[7] Rosales, R.; Sclaroff, S. Inferring body pose without tracking body parts, *CVPR 2000*, pp. 721-727.

[8] G. Shakhnarovich, P. Viola, and T. Darrell. Fast Pose Estimation with Parameter Sensitive Hashing, *ICCV 2003*, pp. 750-757.

[9] C. Sminchisescu, B. Triggs. Covariance Scaled Sampling for Monocular 3D Body Tracking, *CVPR 2001*, 447-454.

[10] C. Sminchisescu, B. Triggs. Kinematic Jump Processes for Monocular Human Tracking, *CVPR 2003*, pp. 69-76.

[11] C.J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU 80(3)*: 349–363, December 2000.

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

[13] P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features, *CVPR 2001*, pp.511-518.

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

[15] T. Zhao, R. Nevatia. Bayesian Human Segmentation in Crowded Situations, *CVPR 2003,* pp.459-466.