Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 761460, 16 pages
doi:10.1155/2010/761460

*Research Article*

# A Two-Stage Bayesian Network Method for 3D Human Pose Estimation from Monocular Image Sequences

**Yuan-Kai Wang and Kuang-You Cheng**

*Department of Electrical Engineering, Fu Jen Catholic University, 24205, Taipei County, Taiwan*

Correspondence should be addressed to Yuan-Kai Wang, ykwang@mails.fju.edu.tw

Received 30 November 2009; Accepted 5 March 2010

Academic Editor: Hsu-Yung Cheng

This paper proposes a novel human motion capture method that locates human body joint position and reconstructs the human pose in 3D space from monocular images. We propose a two-stage framework including 2D and 3D probabilistic graphical models which can solve the occlusion problem for the estimation of human joint positions. The 2D and 3D models adopt directed acyclic structure to avoid error propagation of inference. Image observations corresponding to shape and appearance features of humans are considered as evidence for the inference of 2D joint positions in the 2D model. Both the 2D and 3D models utilize the Expectation Maximization algorithm to learn prior distributions of the models. An annealed Gibbs sampling method is proposed for the two-stage method to inference the maximum posteriori distributions of joint positions. The annealing process can efficiently explore the mode of distributions and find solutions in high-dimensional space. Experiments are conducted on the HumanEva dataset with image sequences of walking motion, which has challenges of occlusion and loss of image observations. Experimental results show that the proposed two-stage approach can efficiently estimate more accurate human poses.

## 1. Introduction

Human pose reconstruction, also called human motion capture, is one of the most popular research topics in computer vision and machine learning. Reconstructing human poses can be used to analyze and recognize human behaviors in many image understanding applications, including human-computer interaction, human-robot interaction, and visual surveillance. Traditionally, motion capture systems use (electro-magnetic or color blob) markers attached to human body to obtain human poses. However, it is uncomfortable and restricts human's freedom to move. In addition, high cost and obtrusion are two major drawbacks of these systems.

Recently, researchers focus on markerless approach which adopts computer vision to analyze human 2D or 3D human poses according to the observations from images [1]. It has the potentiality to provide an inexpensive, non-obtrusive solution and non-restrictive user environment for the estimation of human poses. The complexity of 3D human pose estimation is higher than that of 2D human pose estimation, because 3D human poses constitute a high-dimensional search space for the computation. Some papers utilize multiple cameras to reconstruct 3D human poses due to that depth information can be obtained to ease the problem. However, reconstructing 3D human poses from monocular image sequences is more attractive because it has many advantages such as convenient to use, less restrictions and no computational loading of camera calibration. The challenge of this approach is to infer 3D poses of a highly articulated, self-occluding human gesture from 2D motion sequences without depth information. It is an ill-posed problem because different 3D human poses can have the same 2D projection in the image space due to the occlusion problem of human body such as hidden hands and lower limbs occluded by other body parts.

Model-free (or discriminative) and model-based (or generative) methods are two kinds of approaches for markerless monocular human motion capture. Different from model-free method that searches approximate human pose, model-based method provides more exact estimation result of human pose details. Bayesian network is one of efficient ways in model-based approaches. Bayesian network approach can model the kinematics of body configuration with probabilistic graphical networks. It constitutes a function mapping from features in images to 3D poses of human body. The estimation of 3D poses is achieved by inferencing posteriori distribution of human body parts from belief propagation in the probabilistic graphical networks. Although Bayesian network may solve the occlusion problem, an efficient inference algorithm has to be discovered.

Previous works inference 3D human poses from monocular image sequences directly by image observations. However, human motions with high-degree freedom usually cause self-occlusion and unpredictability. The function mapping from image observations to 3D poses is highly complex.

Rather than the classical algorithms, in this paper, the function mapping is decomposed into two stages. The two-stage model includes a 2D pose and a 3D pose belief propagation networks. The first stage will estimate 2D poses from image observations, and the second stage reconstructs 3D poses from the estimated 2D results. The adding of 2D human pose information can decompose the indirect causal relationship between image observations and 3D human poses. More accurate inference results can be obtained by reducing the computational complexity of the 3D pose estimation problem. Especially, an annealed Gibbs sampling method is proposed for the inference of the two-stage model. The inference algorithm adopts an annealing process for the sampling of posteriori distributions and can efficiently reconstruct 3D poses. The proposed algorithm has lower cost in computing time than that of Gibbs sampling methods.

Figure 1 sketches the proposed two-stage pose estimation method. The method includes three important steps that are feature extraction, 2D pose estimation, and 3D pose estimation. The feature extraction step obtains image observations, such as silhouette and appearance features of human. The 2D pose estimation step inferences 2D human poses via these image observations by an annealed Gibbs sampling algorithm. The result of 2D pose estimation step is used to be the observations of the 3D pose estimation step. Both the 2D and 3D models are trained by the Expectation Maximization (EM) algorithm before estimation process. The training features in 2D model training block are the visual features obtained by the feature extraction block. The training features adopted in the 3D model training block include 2D inference results and parts of visual features that can provide 3D information.

The remainder of this paper is structured as follows. Section 2 reviews earlier related work. Our approach is presented in Sections 3 and 4. In Section 3, a formal description of the proposed two-stage probabilistic framework is presented. This is followed in Section 4 by the extraction of image observations of the human poses. Experimental results are reported in Section 5 to discuss the performance of the approach. Section 6 concludes the paper.

## 2. Review of Previous Work

Model-free approach [2–7] is one kind of monocular pose estimation approach that establishes a direct relation between image observation and human motion. It recovers the kinematics configuration of a person whose appearance in images varies due to different clothing and lighting conditions. This approach exploits model variations in pose configuration, human shape, and appearance, without assuming human body model. Agarwal and Triggs [8] proposed an algorithm to estimate 3D human model from silhouette of single view using nonlinear regression to model the relation between histograms of shape contexts and human pose. They [9] also used histograms of gradient orientations over a grid of small cells with nonnegative matrix factorization to gain a set of basis vectors corresponding to local features of body parts. Bowden et al. [10] extracted 2D silhouette of a moving human and the corresponding 3D skeletal structure to be image observations and adopted a non-linear point distribution model (PDM) to fit them. Silhouette is a good feature to help recover 3D poses since it is insensitive to the variations in color and texture [9]. Rohr [11] used edge lines to replace edges in order to partially get rid of silhouette noise. Although all these model-free methods provide interesting results, the robustness of these methods will suffer since articulated relationship of human body parts is not incorporated to solve the occlusion problem.

The goal of model-based approach is to construct the function that gives the likelihood of the image, given a set of parameters which include body configuration parameters, body shape and appearances. The approach usually adopts an articulated model to represent the relationship among human body parts as a kinematics tree, consisting of divisions linking by joints. The prior knowledge describing kinematic properties by shape, texture and appearance has been used to make the problem tractable. One of the fundamental representations is the Johansson's moving light displays [12]. It adopted a relatively simple representation of the human body called the stick figure that consists of line segments linked by joints. It was demonstrated that relatively little information (motion of a set of selected points on the body) is needed for humans to perform reconstruction of the body configurations. Pose estimation is challenging due to its non-rigid nature, self-occlusion, variable appearance, and high degree of freedom. By incorporating articulation knowledge, model-based approaches are able to overcome these problems to a great extent and are actively explored by many researchers.

In decades, many papers addressed that a probabilistic stick figure represented by belief propagation networks can successfully recover human pose configurations. Belief propagation network has two main subclassifications: Markov network and Bayesian network. Markov networks organized by undirected acyclic graph are probabilistic propagation models that the linking edge between two nodes has no direction. Every node obtains probabilistic feedback from their neighbors. Human models defined by Markov networks consider that body parts and joints are estimated from their neighbor parts [13, 14] estimated human motion and poses

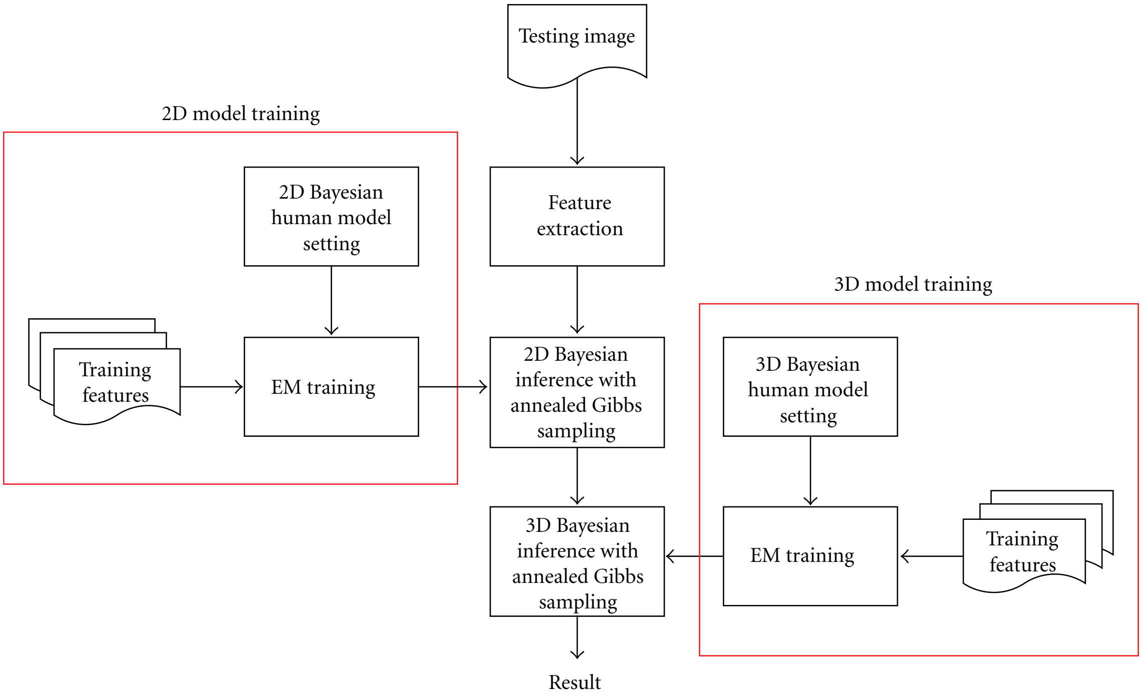

Testing image
2D model training
2D Bayesian human model setting
Feature extraction
3D model training
Training features
EM training
2D Bayesian inference with annealed Gibbs sampling
3D Bayesian human model setting
3D Bayesian inference with annealed Gibbs sampling
EM training
Training features
Result

Figure 1: Flowchart of the proposed method.

using Nonparametric Belief Propagation and particle filters. Hua et al. [15] exploited a data driven belief propagation Monte Carlo algorithm with importance sampling functions building from bottom-up visual cues for efficient inference. Using Markov network has the advantage that every node would influence their neighbors and converges to a stable result. However, errors may be easily propagated to all nodes in the network.

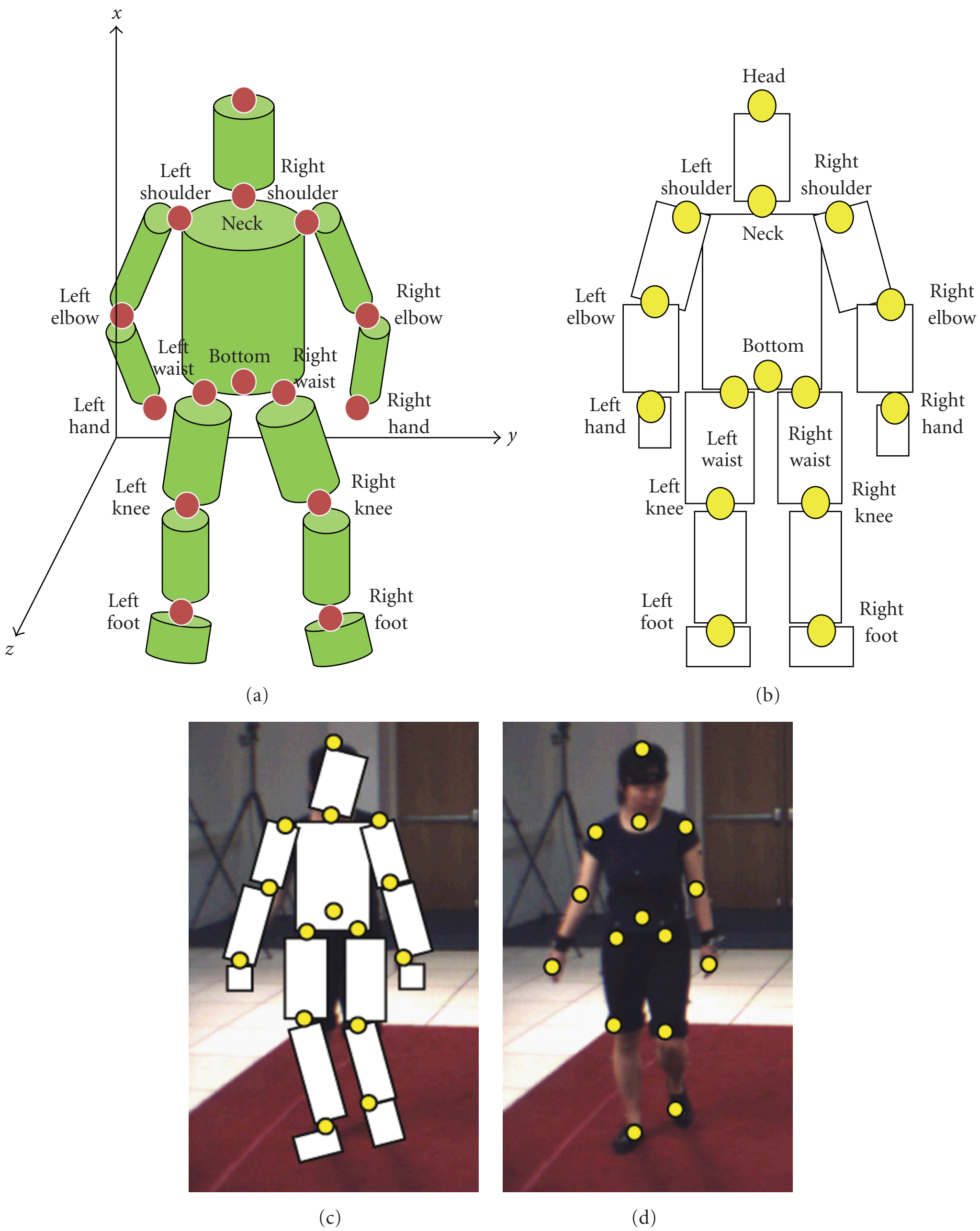
In Bayesian network models, nodes have directed arcs linked to other nodes, which are also called a directed acyclic graph (DAG). These arcs represent the relation of cause and effect between nodes. Parent nodes in DAG are not influenced by their child nodes. Human body parts with stable movement, such as torso, are usually considered to be the root node and/or parent node. Child nodes represent the body parts with high freedom of movement that have large location variations and can also represent image features [16–18]. Leonid and Darrell [19] used a generative appearance model defined by Bayesian networks. The Bayesian generative appearance model is similar to [20] for 3D articulated tracking and [21] for modeling the interactions of multiple independent moving objects. Compared with Markov networks, Bayesian network is more efficient when there are stable visual features assistant for the finding of body parts. Another advantage is that estimation errors of specific body objects would not propagate to other body elements.

Previous works of belief propagation approach pay attention on one-stage framework, which estimates 3D human poses directly from image observations. However, obtaining 3D body pose from visual features can be considered as a machine learning problem that approximates a function mapping of a given data space to targeted pose space. This function seems to be highly complex. Since same visual features can represent different body pose configurations and same body configurations can generate different visual features due to clothing and view-point, it makes the ill-posed problem more difficult. Therefore, the search of posterior distributions in high-dimensional state space from sparse learning data is still intractable. Motivated partly by these issues, a two-stage framework with an annealed Gibbs sampling inference algorithm which can increase estimation accuracy and reduce computation time is proposed in this paper.

## 3. The Two-Stage Framework for Pose Estimation

This section will give a two-stage Bayesian network framework including 2D and 3D articulated human models for the estimation of human poses. The proposed Bayesian network is a belief propagation network using an annealed Gibbs sampling algorithm to estimate 2D and 3D human joint positions.

*3.1. The Probabilistic Graphical Networks for Articulated Human Modeling.* The articulated human model is defined by a set of reference points. The 3D and 2D articulated human models consisting of 15 distinctive points are shown in Figures 2(a) and 2(b). Those points include a reference point of head and 14 joint points. Figure 2(c) overlays a particular 2D model on the human in the corresponding image. Figure 2(d) illustrates all 2D joint positions we need to estimate.



(a)

(b)

(c)

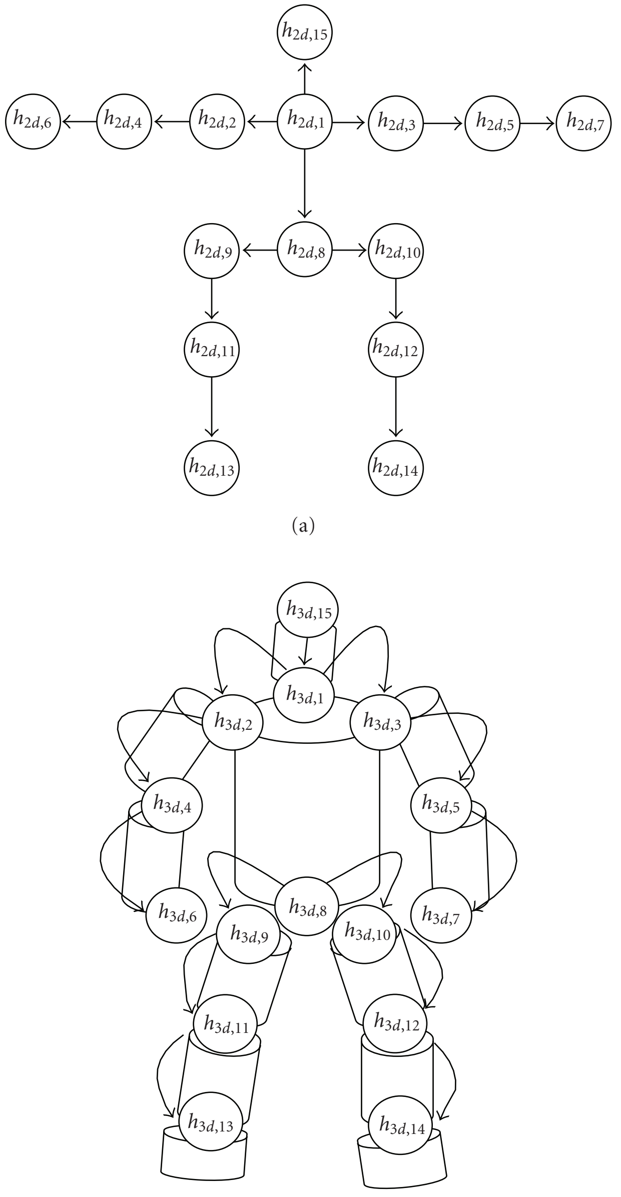
(d)

Figure 2: Articulated human model. (a) 3D model and joints, (b) 2D model and joints, (c) projected model on the image, (d) joint positions mapping to the human body.

Directed graph is used for the abstraction of the articulated human models. The human pose is defined as a node set $H = \{H_{2D}, H_{3D}\}$, where $H_{2D} = \{h_{2d,1}, \dots, h_{2d,15}\}$, $H_{3D} = \{h_{3d,1}, \dots, h_{3d,15}\}$, and $h_{2d,i}$ and $h_{3d,i}$ are the 2D and 3D positions of the $i$th joint. Shown in Figures 3(a) and 3(b) are the graphical model of humans. According to the directed graphical model, each node is affected only by its parent node.

The representation of 2D and 3D poses is a probabilistic stick figure with joint positions but not joint angles. The disadvantage of joint angle representation is that more body parameters such as limb length and global orientation of body have to be given for pose estimation. On the other hand, joint position can uniquely encode the pose of human.

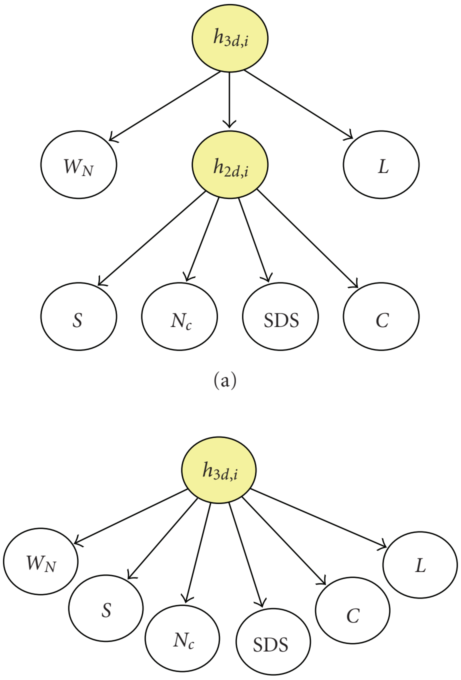
A Bayesian network is a directed acyclic graph $\vec{G} = (V, \vec{E}, C)$ defined by a set of nodes $V$, a set of directed edges $\vec{E}$, and an edge cost function $C$ defined as $C : (i, j) \rightarrow R^{+}$ for all $(i, j) \in \vec{E}$. To encode probabilistic information into the graph, each node is regarded as a random variable. An edge indicates a probabilistic dependency between the parent node and the child node. Besides the graph structure, $C$ is considered as a set of conditional probability distributions governing the probability of a particular value of a node



(a)

(b)

Figure 3: Graphical model. (a) 2D graphical model, (b) 3D graphical model.



(a)

(b)

Figure 4: Comparison of one-stage and two-stage networks for pose estimation. (a) The proposed hierarchy of causal relationship among observations, 2D poses, and 3D poses, (b) one-stage network directly linking observations and 3D poses.

given an instantiation of its parents. In other words, for each node $V_i$ with parents $V_j \in pa(V_i)$, an edge $(i, j)$ exists and there is a conditional probability distribution $P(V_i \mid V_j)$ representing the edge cost function for the edge; for each $V_i$ without parents, there is a prior probability distribution $P(V_i)$.

We denote our 2D and 3D Bayesian networks as $\vec{G}_{2D} = (V_{2D}, \vec{E}_{2D}, C_{2D})$ and $\vec{G}_{3D} = (V_{3D}, \vec{E}_{3D}, C_{3D})$, where $V_{2D} = \{H_{2D}, O_{2D}\}$ and $V_{3D} = \{H_{3D}, O_{3D}\}$ are two sets of random variables describing the joints and observations of human body parts. The 2D observation set $O_{2D} = \{f_k, 1 \leq k \leq n\}$ consists of the visual features: contour point set, normalized center, spatial distribution of skin color, and corners, which will be discussed in Section 4. The 3D observation set $O_{3D} = \{O_{3d,1}, \ldots, O_{3d,15}\}$ is defined by $O_{3d,j} = \{h_{2d,j}, w_N, L\}$ includes not only some visual features, but also the corresponding 2D joint positions of the body parts $h_{2d,j}$ as the observations. The directed edges $\vec{E}_{2D} = \{(h_{2d,i}, h_{2d,j}), (h_{2d,i}, O_{2D})\}$ and $\vec{E}_{3D} = \{(h_{3d,i}, h_{3d,j}), (h_{3d,i}, O_{3d,i})\}$ can be categorized into two edge sets. The edges $(h_{2d,i}, h_{2d,j})$ and $(h_{3d,i}, h_{3d,j})$ link joint nodes and are illustrated in Figure 3. The edges $(h_{2d,i}, O_{2D})$ and $(h_{3d,i}, O_{3d,i})$ link joint nodes and observation nodes and are illustrated in Figure 4(a). It is the proposed two-stage hierarchy which is different with usual one-stage networks shown in Figure 4(b). The edge cost functions of 2D and 3D models, $C_{2D} = \{P(h_{2d,i} \mid pa(h_{2d,i}))\}$ and $C_{3D} = \{P(h_{3d,i} \mid pa(h_{3d,i}))\}$, are conditional distributions obtained by EM learning algorithm. The proposed two-stage hierarchy decomposes the direct function mapping of visual features to 3D poses into two steps: first mapping visual features to 2D poses, and then mapping 2D poses and visual features to 3D poses. The method can increase accuracy since the exploration space in each stage is smaller than that in the direct function mapping.

The proposed DAG encodes the cause-effect and conditional independence relationships among variables into a probabilistic reasoning system. The directions of links represent causality. The links between the nodes, or variables, represent the conditional probabilities of inferring the existence of one variable given the existence of the other variable.

The proposed network structures specify two unique joint probability distributions (JPD) $P_{2D}(V_{2D})$ and $P_{3D}(V_{3D})$. They will be simply denoted as $P(V)$ in the following derivations since both distributions have the same Markov property. $P(V)$ reflects the properties of the network and can be factorized into the product of all conditional probability distributions

$$P(V) = \prod_{i=1}^{n} P(V_i \mid pa(V_i)), \tag{1}$$

where $n$ is the number of all nodes in the network. This factorization of the JPD comes from a local Markov property called *Markov blanket* [22]. Namely that each variable is independent of its nondescendents in the graph given the state of its parents. The *conditional independence* property can reduce, sometimes significantly, the number of parameters that are required to characterize the JPD of the variables. This reduction provides an efficient way to compute the posterior probabilities given the evidence.

There are two problems for the computation of human poses $H$: the parameter learning of the conditional distributions $P(V_i \mid pa(V_i))$, and the inference of the posterior probabilities of $H$ given the evidence $O$, $P(H \mid O)$. Details will be given in next two subsections.

*3.2. Parameter Estimation by EM Algorithm.* Learning process is important for Bayesian network. The training data consisting of human poses and observations is incomplete and sparse. The learning algorithm of Bayesian network has to be built on a high-dimensional solution space and solve the local maxima problem. In this paper, we use the EM algorithm [23] to train the 2D and 3D graphical models.

In the proposed Bayesian network, the local probability distribution for each node is formulated by $P(V_i \mid pa(V_i), \theta, S^h)$, where $\theta$ is the parameter vector of probability distribution, and $S^h$ is the topology of the 2D or 3D Bayesian network. Parameter learning is used to find a good approximation parameter $\widehat{\theta}$ for $\theta$ that can be explained by the training data set $D = \{D_1, \dots, D_N\}$, where $N$ represents the number of training samples. $D_l = \{V_1[l], \dots, V_n[l]\}$ is the $l$th training sample. A log-likelihood function $L_D(\theta) = \log(P(D \mid \theta))$ is formulated based on the independence assumption of training samples and can be decomposed into the product of all conditional distributions according to the conditional independence property:

$$\begin{aligned} L_D(\theta) &= \log\left\{\prod_{l=1}^{N} P(V_1[l], \dots, V_n[l] \mid \theta)\right\} \\ &= \sum_{i=1}^{n}\sum_{l=1}^{N} \log P(V_i[l] \mid pa_i(V_i(l)), \theta). \end{aligned} \tag{2}$$

We obtain the parameter $\widehat{\theta}$ by the Maximum Likelihood Estimation (MLE) method: $\widehat{\theta} = \arg\max_{\theta} L_D(\theta)0$.

However, the training data $D$ is incomplete because of self-occlusion of body parts. That is, the variables $V_i[l]$ or parents of the variable $V_j[l] \in pa(V_i[l])$ in (2) are hidden and missing. The partial observability induces the so-called incomplete data problem. Hence, MLE can not be directly utilized for the training. The EM algorithm is a common method solving the learning of incomplete data.

Assume $D$ is incomplete. We can claim $D = Y \cup U$, where $Y$ is an observed data set and $U$ is the missing data set. The EM learning algorithm guesses a probable initial $\theta^{(0)}$ at the start of learning process, and then proceeds to repeatedly generate $\theta^{(t)}$ iteratively by Expectation-step and Maximization-step.

The Expectation-step computes the conditional expectation of the log likelihood function with a given $Y$ and the current parameter $\theta^{(t)}$:

$$Q\left(\theta \mid \theta^{(t)}\right) = E_{\theta^{(t)}} = \left[\log P(D \mid \theta) \mid \theta^{(t)}, Y\right]. \tag{3}$$

The Maximization-step updates the $t+1$ step parameter $\theta^{(t+1)}$ from current parameter $\theta^t$ under the assumption that the Expectation-step computes a correct $\theta^t$:

$$\theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta \mid \theta^{(t)}\right). \tag{4}$$

The Expectation-step and Maximization-step are repeated until the difference of $L_D(\theta^{(t+1)}) - L_D(\theta^{(t)})$ converges.

*3.3. Posterior Inference by Annealed Gibbs Sampling.* Estimating human poses given a set of observations can be formulated as a probabilistic inference problem. Let the observed data be $O' = O - U$, where $U$ is the set of hidden variables that are unobservable due to occlusion. The best estimated pose is a vector $H^*$, which is defined as the pose with the maximum probability given $O'$. The posterior probability $P(H \mid O')$ is maximized over all possible pose space according to Bayes' rule and marginalization as follows:

$$\begin{aligned} H^* &= \arg\max P(H \mid O') \\ &= \arg\max \int_{u \in U} P(H, u \mid O')du \\ &= \arg\max \int_{u \in U} P(H, O', u)du, \end{aligned} \tag{5}$$

with the assumption that the prior distribution $P(H)$ is uniform. The function $P(H, O', u)$ is a specific case of the full JPD $P(V)$ because $V = H \cup O' \cup U$. Substituting (1) into (5) we obtain the inference formulation:

$$H^* = \arg\max \int_{u \in U} \prod_{i=1,\dots,n} P(V_i \mid pa(V_i))du. \tag{6}$$

It states that given the Bayesian network specified the JPD in a factored form, we can estimate poses by evaluating all possible inference queries by marginalization, that is, integrating out over hidden variables.

The direct way for the full integration over continuous variables is called exact inference that can obtain precise inference results when querying the Bayesian network. Although the high-dimensional JPD is factored into the product of low-dimensional conditionals, it still takes high computational complexity and known to be an NP-hard problem. Some algorithms, such as junction tree and message passing algorithms, can efficiently inference $H^*$ only in restricted classes of networks.

Approximate inference methods [24] have also been proposed in the literature, such as the loopy belief propagation, variational methods, and Markov chain Monte Carlo (MCMC) sampling methods. Approximate inference methods have low computational complexity. MCMC [25] is attractive because it solves the integration problem in high-dimensional space by gradually improving estimates as sampling proceeds. A variety of standard MCMC methods, including the *Metropolis-Hastings algorithm* and the *Gibbs sampling*, were used for approximate inference.

The MCMC sampling methods obtain samples of the target distribution $P(V)$ by Markov chain. A sample vector $v$ is defined as a sample of the random vector $V$. At $t$ iteration, a candidate vector $v^*$ is sampled from $v^{(t-1)}$, the sample vector of $V$ obtained at $t-1$ iteration, and a proposal distribution $q(v^* \mid v^{(t-1)})$. The candidate is accepted as the new state $v^{(t)}$ with a probability $\alpha(v^{(t-1)}, v^*) = \min(1, p(v^*)q(v^{(t-1)} \mid v^*)/p(v^{(t-1)})q(v^* \mid v^{(t-1)}))$. This generates a Markov chain $(v^{(0)}, v^{(1)}, \ldots, v^{(k)}, \ldots)$, as the transition probabilities from $v^{(t-1)}$ to $v^{(t)}$ depends only on $v^{(t-1)}$ and not $(v^{(0)}, v^{(1)}, \ldots, v^{(t-2)})$. Following a sufficient burn-in period (of, say, $k$ steps), the chain approaches its stationary distribution and samples from the vector $(v^{(k+1)}, \ldots, v^{(k+n)})$ are samples from $P(V)$. The Monte Carlo integration method can then be applied to solve (5). However, the sampling of $P(V)$ is still not efficient if $V$ is in a high-dimensional space.

In this paper, a Gibbs sampling method with annealing process is proposed for the inference of the Bayesian network. The key to the annealed Gibbs (AG) sampler is that we only consider univariate conditional distributions sampled with a stochastic process controlled by simulated cooling process. The AG sampler is first proposed by S. Geman and D. Geman [26] for Markov random fields with Gibbs field. Here the sampler is revised for the proposed two-stage Bayesian network.

We define an expression of full conditionals as $P(V_j \mid V_{-j}) = P(V_j \mid V_1, \ldots, V_{j-1}, V_{j+1}, \ldots, V_n)$ where $V_{-j}$ denotes a vector containing all of the variables but $V_j$. It follows by the Markov blanket property [25] that the full conditionals can be simplified as follows: $P(V_j \mid V_{-j}) = P(V_j \mid pa(V_j))\prod_{k\in \mathrm{ch}(V_j)} P(V_k \mid pa(V_k))$, where $\mathrm{ch}(V_j)$ denotes the children nodes of $V_j$. That is, we only need to take into account the parents, the children, and the children's parents. Sampling from the full conditionals, with the AG sampler, lends itself naturally to the construction of general purpose MCMC.

We define the samples of $V_j$, the $j$th variable of $V$, as $v_j$. The samples $v_j$ are drawn from the distribution $P(V_j \mid V_{-j})$. Thus, to obtain the $v_j^{(t)}$, that is, $v_j$ in the $t$th iteration of the sampling process, we draw from the distribution

$$v_j^{(t)} \sim P\left(V_j \mid v_1^{(t)}, \ldots, v_{j-1}^{(t)}, v_{j+1}^{(t)}, \ldots, v_n^{(t)}\right). \qquad (7)$$

Therefore, the AG sampler adopts the proposal distribution for $j = 1, \ldots, n$

$$q\left(v^* \mid v^{(t)}\right) = \begin{cases} p\left(v_j^* \mid v_{-j}^{(i)}\right) & \text{if } v^*_{-j} = v^{(t)}_{-j} \\ 0 & \text{otherwise.} \end{cases} \qquad (8)$$

If the probability of the move for $v_j^{(t)}$ is $\alpha(v_j^{(t)}, v^*) = \min(1, p(v^*)q(v_j^{(t)} \mid v^*)/p(v_j^{(t)})q(v^* \mid v_j^{(t)})) = 1$, it becomes a Gibbs sampler. However, the AG sampler is proposed to improve the iteration process. It is similar to Metropolis sampling but has different probability $\alpha$ of a move given by

$$\alpha_{\mathrm{AG}} = \min\left(1, \left(\frac{p(v^*)}{p\left(v_j^{(t)}\right)}\right)^{1/T(t)} \frac{q\left(v_j^{(t)} \mid v^*\right)}{q\left(v^* \mid v_j^{(t)}\right)}\right). \qquad (9)$$

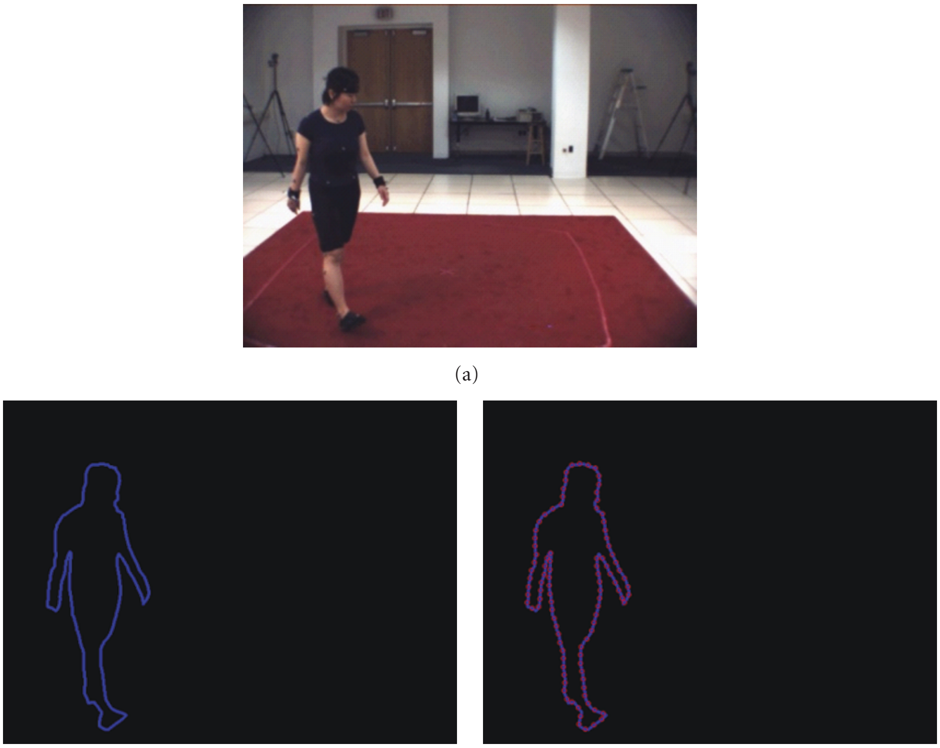
The idea is that when we initially start sampling the space, we will accept a reasonable probability of a down-hill move in order to explore the entire space. As the process proceeds, we decrease the probability of such down-hill moves. Function $T(t)$ is called cooling schedule and the particular value of $T$ at any point in the chain is called the temperature. Typically, $T_0$ is start time and $T_f$ is the final cool down temperatures over $n$ step. A typical setting of geometric decline for the temperature is given by

$$T(t) = T_0\left(\frac{T_f}{T_0}\right)^{t/n}. \qquad (10)$$

The AG sampler adopts a stochastic iterative algorithm that converges to the set of points which are the global maxima of the given function. The annealing process can find the maximum of complex distributions with multiple peaks where standard hill-climbing approaches may trap the algorithm at a less optimal peak.

The advantage of the AG sampler is its efficiency compared to the Gibbs sampler. Instead of wanting to approximate $P(V)$, we want to find the global maximum, that is, the ML estimate of posterior distribution. We could run a Markov chain of invariant distribution $P(V)$ and estimate the global mode by $H^* = \arg\max_{V^{(t)}, i=1,\ldots,N} \int_{u\in U} P(V^{(t)})du$. However, this method is inefficient because the random samples only rarely come from the vicinity of the mode. Unless the distribution has large probability mass around the mode, computing resources will be wasted exploring areas of no interest.

This annealing process involves simulating a nonhomogeneous Markov chain whose invariant distribution at iteration $t$ is no longer equal to $P(V)$, but to $P_t(V) \propto P^{1/T(t)}(V)$, where $\lim_{t\to\infty} T(t) = 0$. The reason for doing this is that, under weak regularity assumptions on $P(V)$, $P^{\infty}(V)$ is a probability density that concentrates itself on the set of global maxima of $P(V)$. The simulated annealing involves, therefore, just a minor modification of standard MCMC algorithms.

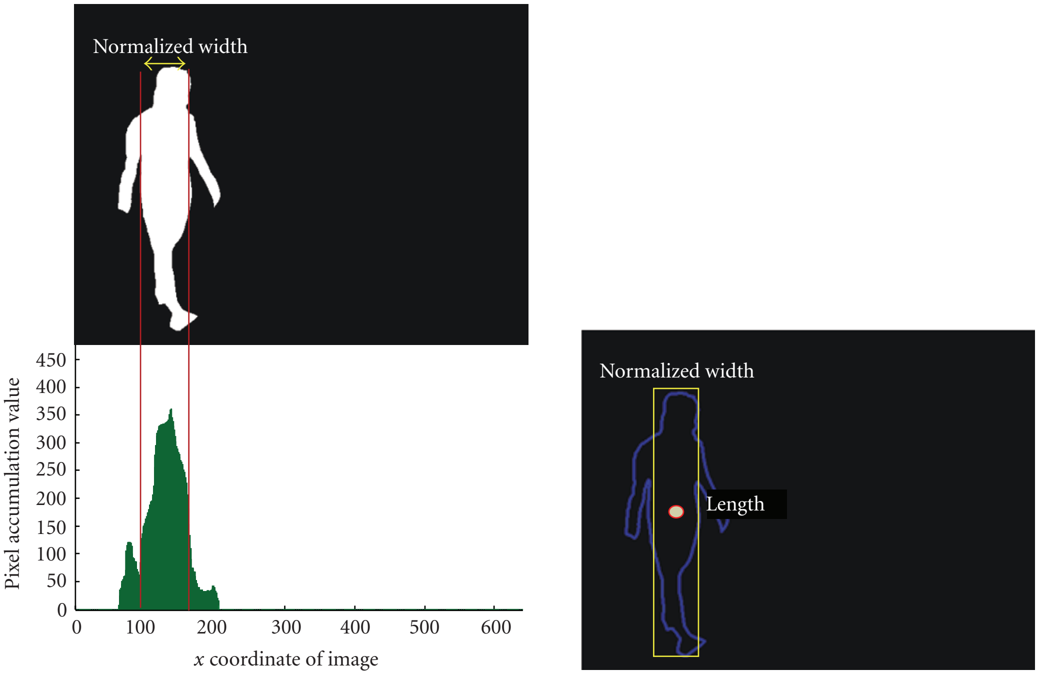


(a)

(b)

(c)

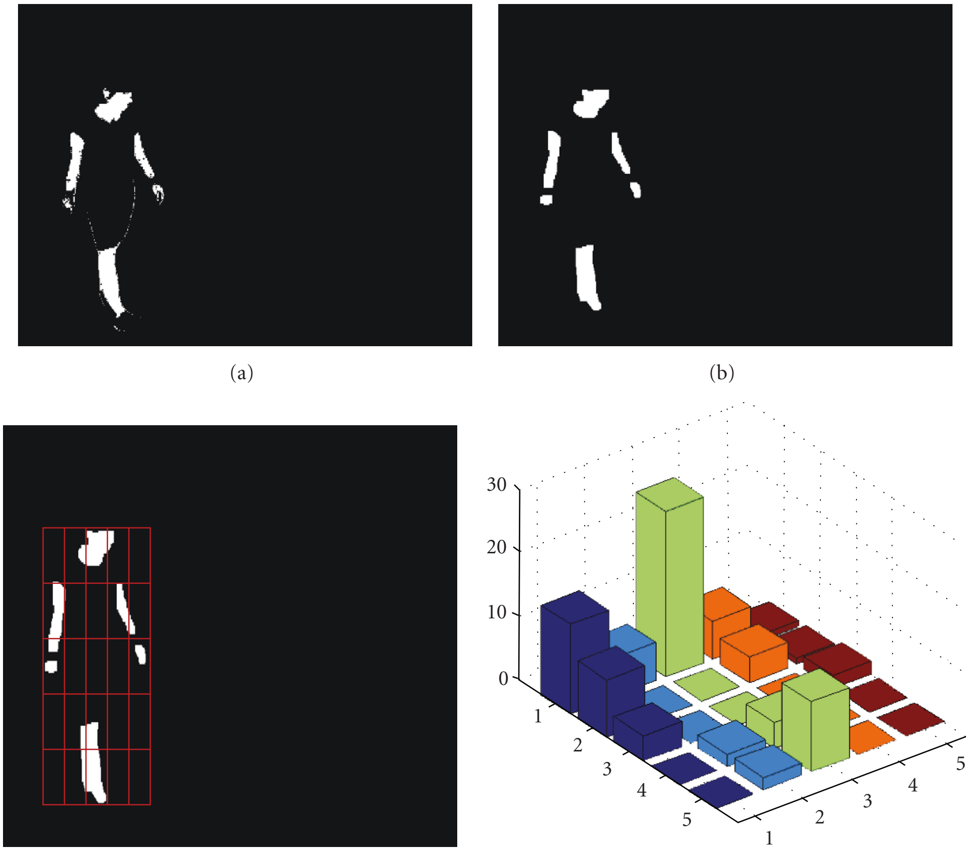
Figure 5: Human silhouette. (a) Original image, (b) extracted contour of the human, (c) uniformly sampled points of the extracted contour.



(a)

(b)

Figure 6: Normalized width of human shape.



(a)

(b)

(c)

(d)

Figure 7: Spatial distribution of skin color. (a) Skin color detection, (b) noise removal and pixel grouping by morphology, (c) region segment, (d) the spatial distribution of skin color in (c).

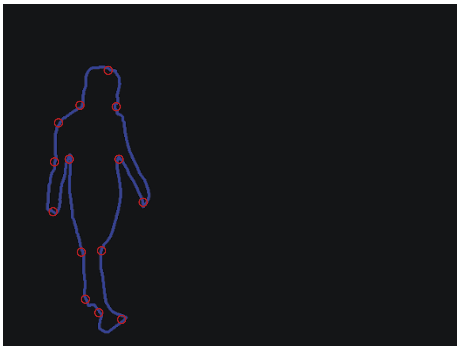


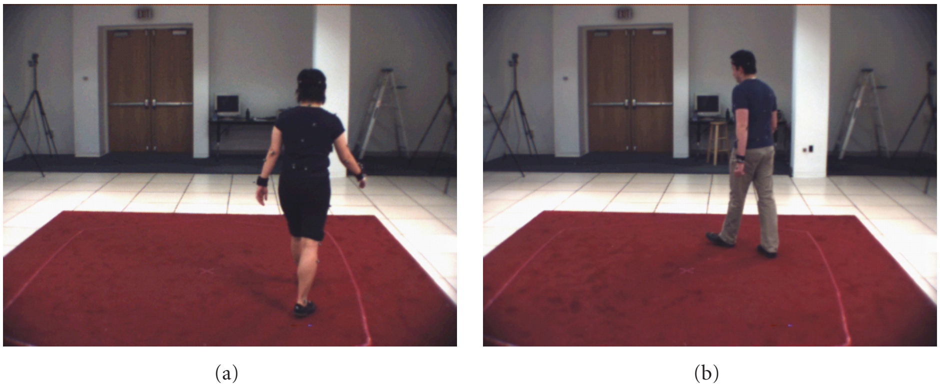
Figure 8: Corners of silhouette.

## 4. Feature Extraction

Image features, such as edge, color, and silhouette, are observations of a pose. The extraction of image features constitutes evidence nodes of the articulated human model for the inference in the Bayesian network. Single human feature is not enough to inference 3D human position since different 3D poses can exhibit similar 2D observations in the images. Therefore, we devise 4 kinds of features in the proposed method: human silhouette, normalized center of human body, spatial distribution of skin color, and corners of human body.

*4.1. Human Silhouette.* Human silhouette is a useful feature with highly resistance to the change of appearance and illumination. Silhouette features contain full body information, including head, torso, and limbs. Figure 5(a) is an original image from HumanEva I database [27, 28]. The subject image is segmented from background and transformed to a binary image. Figure 5(b) is the contour image extracted by topological structural analysis. The extracted contour is then uniformly sampled into a contour point set, which is depicted in Figure 5(c).

*4.2. Normalized Center.* Normalized center is proposed to normalize position-dependent features, such as silhouette. However it cannot be readily obtained because human's shape varies frame by frame due to its nonrigid characteristics caused by limbs motion. A normalized width $w_N$, instead of width of subject contour boundary, has to be calculated. We use the normalized width to reduce the effect of limbs motion. As shown in Figure 6(a), we calculate the profile along $x$ coordinate and denote the accumulation value in the profile as $h_x$. The $w_N$ is defined as $w_N = x_R - x_L$, where $x_L = x$



(a)

(b)

Figure 9: Sample images of the walking sequences used in our experiments.

if $h_x \geqq$ threshold and $h_{x-1} <$ threshold for $x = 1 \rightarrow w$, and $x_R = x$ if $h_x \geqq$ threshold and $h_{x+1} <$ threshold for $x = w \rightarrow 1$. The threshold is determined by empirical study from experimental results.

As we get $w_N$ we can obtain the normalized boundary rectangle shown in Figure 6(b). We calculate the center of normalized boundary rectangle that is called normalized center $N_c$ as shown in Figure 6(b). Assume that the left top point of normalized boundary rectangle is called $p$. We can define the $x$ and $y$ coordinates of the normalized center $N_c$ as $x_N = x_p + 0.5w_N$, $y_N = y_p + 0.5L$, where $L$ is the boundary length of subject. Both $w_n$ and $N_c$ will be used as the observations in the two-stage Bayesian network.

*4.3. Spatial Distribution of Skin Color.* Appearance feature which is proposed in this paper is called spatial distribution of skin color. We use the Gaussian Mixture Model (GMM) to obtain skin color regions [29]. The GMM uses multiple Gaussians to model the probability of skin color pixels in the image. EM algorithm is utilized to find the parameters of the GMM. A 2D histogram of the skin color pixels is then computed to describe to the spatial distribution of those skin color pixels.

Figure 7(a) presents extracted skin color regions of Figure 5(a) by the GMM method. Shown in Figure 7(b) is the result of morphology that is adopted to remove noises and fill in connected components. The minimum bounded area of the subject is then equally divided into several smaller blocks, which are illustrated in Figure 7(c). Accumulated number of skin color pixels in each block is calculated. Figure 7(d) illustrates the ratio of skin color pixels of each block to total skin color pixels in the image, which is the 2D histogram, describing the spatial distribution of skin colors.

*4.4. Corners of Human Body.* Corners of human shape have strong relationship to joints of body parts. It can be exploited by Bayesian networks as good observations for the finding of 2D human body parts. We use the level curve curvature method [30] that defines corners as areas. These areas have level curves multiplied by the gradient magnitude raised to the power of 3 which can intensify the level curves and help the search of local maximum. Figure 8 shows the extracted corners.

The four visual features complement each other. The silhouette feature can well describe the localization of human shape, but is prone to occlusion. Normalized center is helpful for estimating proximal and distal joints of torso. Nevertheless, simply using the feature is not appropriate because the high-degree freedom of limbs influences the boundary's width. Skin color distribution is the feature sensitive to the change of illumination, but offers good indications of the positions of face and limbs. Corners potentially indicate joint locations of human. The locations of knees and feet that we are interested in have high probability close to corners of human silhouette, because they are joints of human.

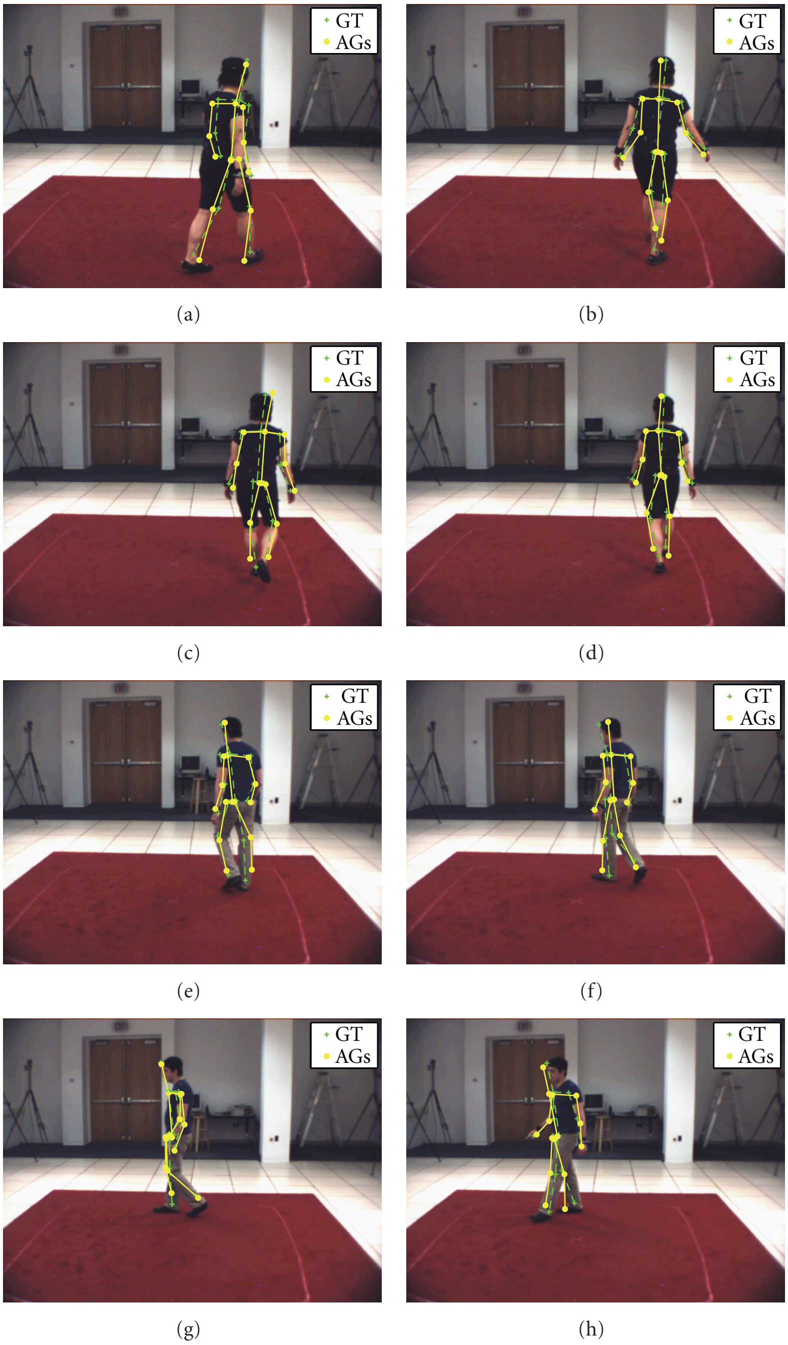
## 5. Experimental Results

*5.1. Experimental Setup.* The proposed method has been evaluated to demonstrate its effectiveness. Experiments are carried out to compare the proposed method with the Gibbs sampling method. The method is implemented with C and C++, and experiments are conducted on a PC with 2.4 GHz Pentium 4 CPU and 1 GB memory.

We presents experimental results conducted on the image sequences in the HumanEva dataset [27, 28]. It is a dataset including both video and ground-truth 3D motion captured by multiple cameras and a calibrated marker-based motion capture system. The image sequences are captured with a resolution of 644 × 488 pixels and a frame rate of 120 Hz. Human activities in the dataset are played by multiple subjects performing a set of predefined actions with a number of repetitions. The dataset has been divided into separate training and testing sets.

The motion capture data given by the HumanEva dataset will be used to train our model and test as ground truth data. The motion capture data provides 3D position information about the location of a set of markers roughly corresponding to a subset of major human body joints. 3D marker positions are projected into 2D marker positions by perspective geometry.

We have run our experiments on the circular walking sequences in the dataset. Subjects walked in an elliptical path



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 10: Results of estimated 2D pose projected onto images. Green dotted lines are the ground truth and yellow lines represent estimation result of our method. (a)–(d) Subject 1. (e)–(h) Subject 2.

three times at the edge of the capture space. The walking sequences contain a full 180° turn that induces the challenges of occlusion of body parts and loss of visual features. Some example images are shown in Figure 9. Note that to demonstrate the method's ability of occlusion handling with monocular images, only the image sequences from single camera are adopted for the experiments. No depth information is utilized. Also, other image sequences such as running and jogging are not tested in our experiments because fast-motion gestures are only critical to tracking problems which is not the concern of this work. The Bayesian network proposed in this paper is used to the detection of reconstructed poses.

*5.2. Performance Evaluation.* This subsection will give comparison results between the proposed two-stage method and classical one-stage method. We first illustrate the results of the two-stage method by super-imposing the estimated 2D poses and ground truth onto the image. Selected frames are presented in Figure 10. It can be seen that our method

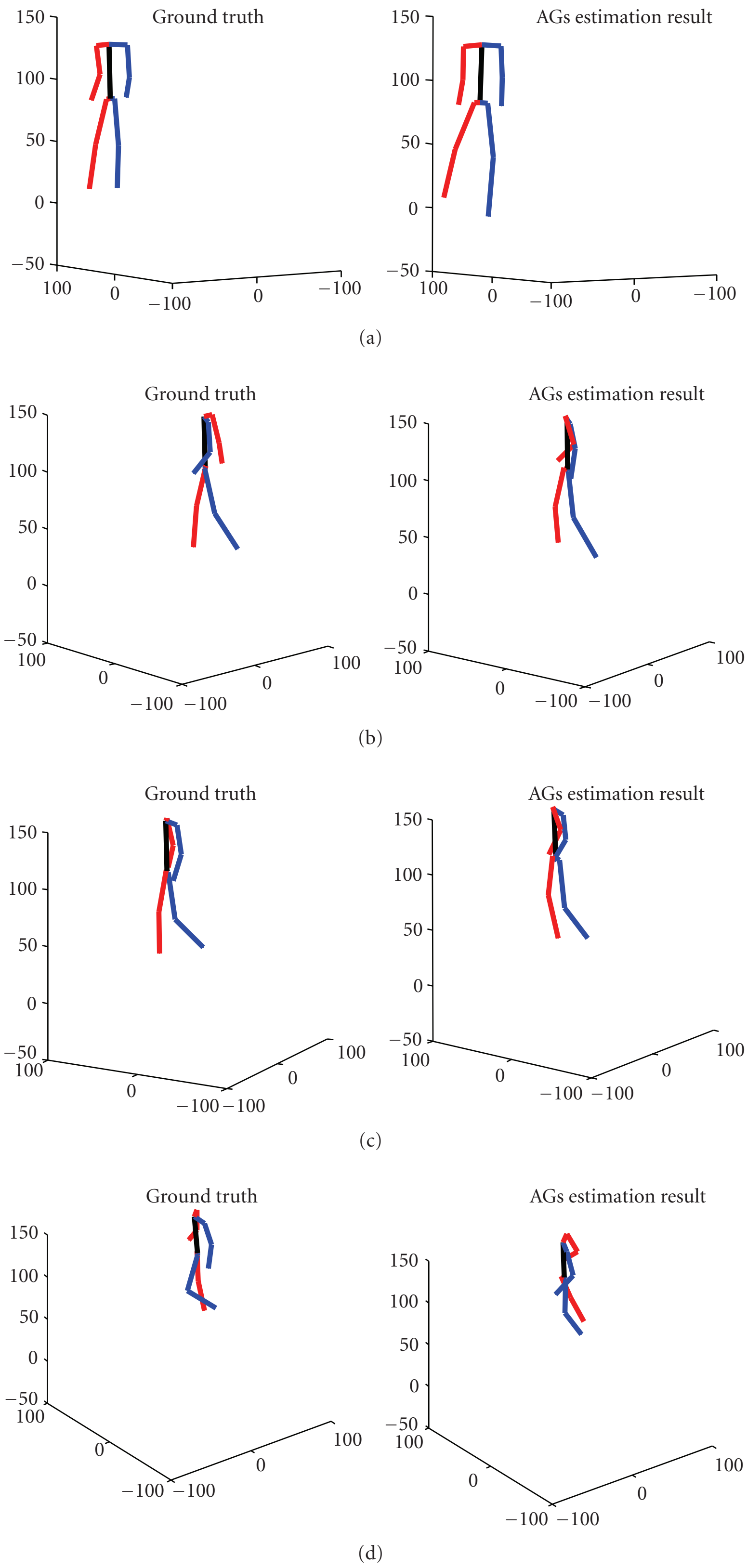
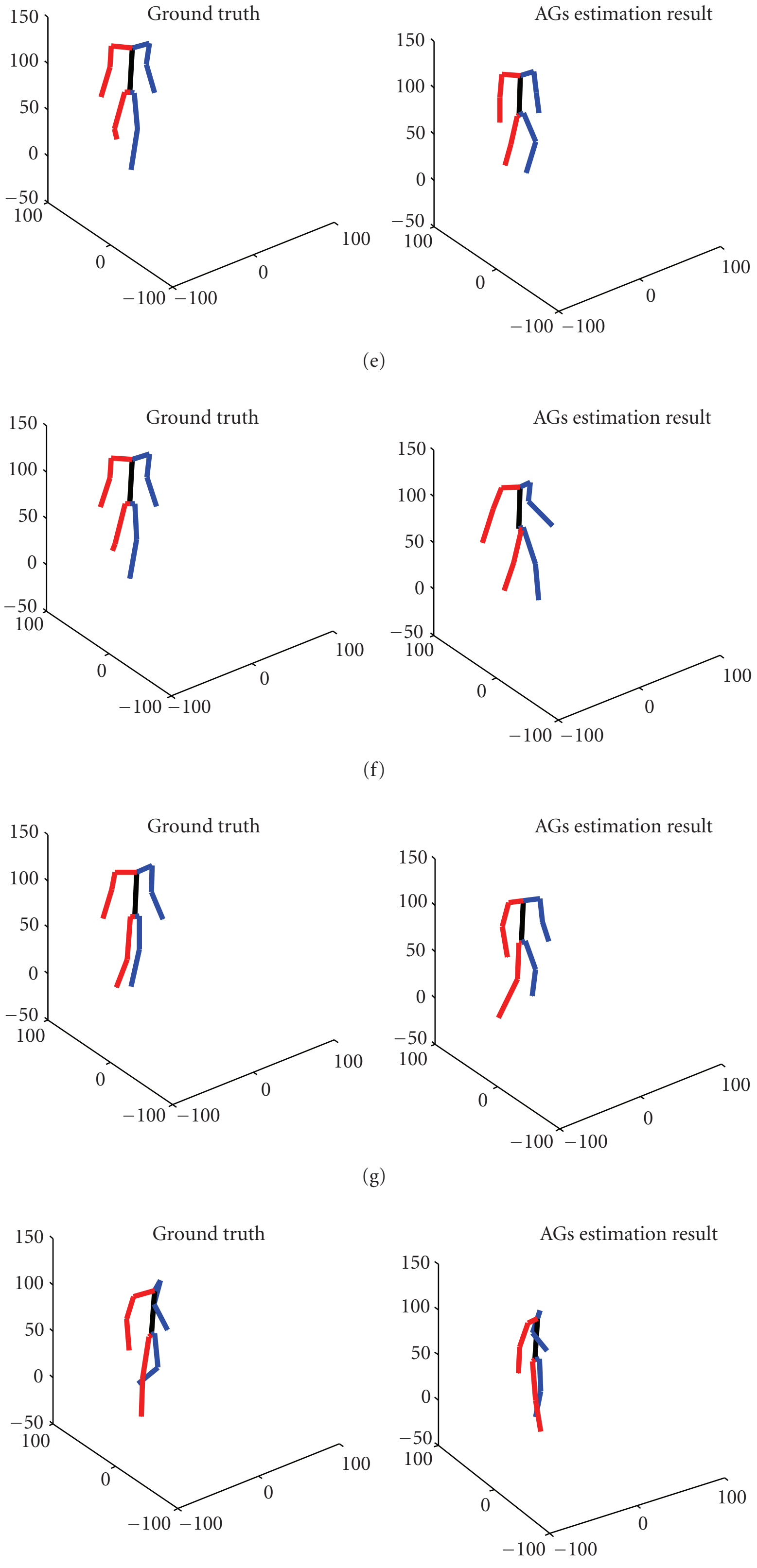


Figure 11: Continued.



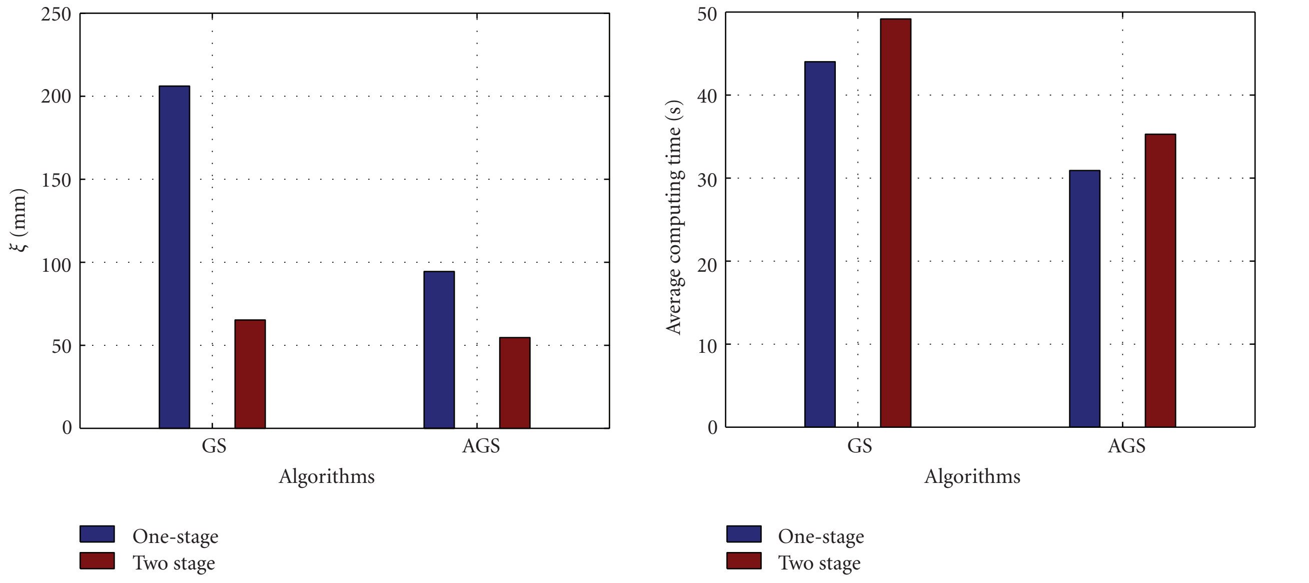
(e)

(f)

(g)

(h)

Figure 11: Results of estimation result in 3D space. (a)–(h) The left figures are ground truth, and the estimated results are shown in the right figure. The units of axes in all figures are centimeter.



(a)

(b)

Figure 12: Performance comparison between the two-stage and one-stage methods. Both the Gibbs sampling and the annealed Gibbs sampling algorithms are implemented for the two methods. (a) Average distance errors, (b) average computing time.

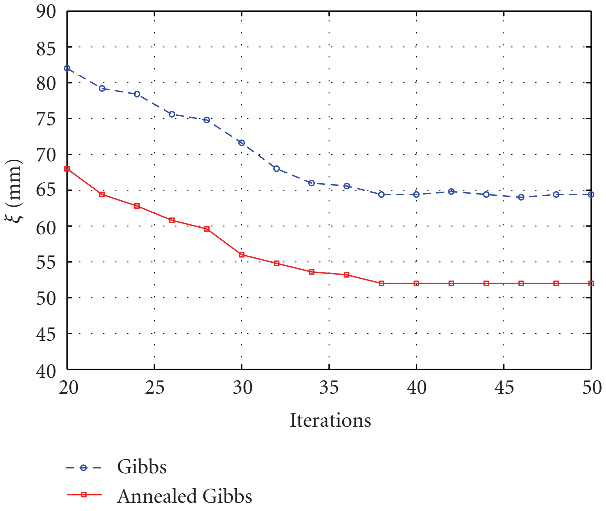


Figure 13: The effect of iteration number on accuracy.

can estimate joint parts that are out of our vision because they are occluded by other body parts. Some local errors, mostly due to the lack of visual features and occluded parts, can be observed. For example, the local error of left arm in Figure 10(a) is produced because the left arm is occluded by torso. The head positions are sometimes less accurate when skin area of face becomes invisible. However, the estimates are still fairly consistent among these results.

Figure 11 presents the estimation results of 3D human structure in 3D space. Our method can simulate the hidden body part to solve the occlusion problem and reconstruct the human pose approximately. On this challenging set of images, the overall 3D pose estimation is fairly satisfactory. It has to be noticed that no tracking algorithm is incorporated in our method. The presented results are obtained frame by frame from monocular images. The local errors can be alleviated by further exploiting pose tracking from multiple cameras, or introducing more accurate feature descriptor.

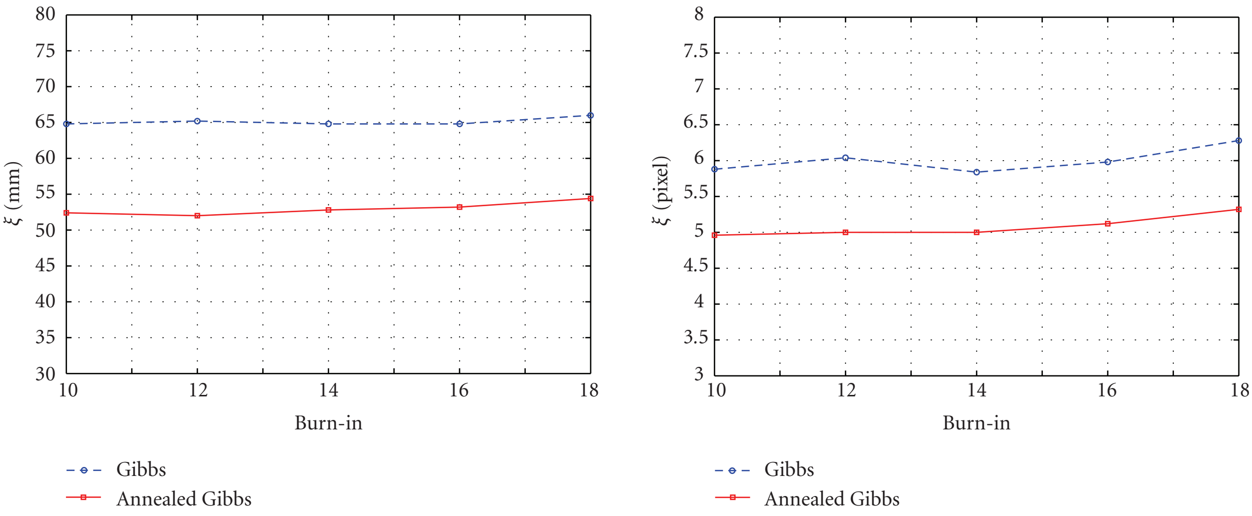
To quantitatively examine the performance of the proposed method, a measurement is defined by a distance error of poses between estimated results and ground truth. Suppose $H = \{h_1, h_2, \ldots, h_M\}$, where $h_m \in R^3$ (or $x_m \in R^2$ for the 2D body model), is the position vector of the body pose in the world (or image, resp.). The error in estimated pose $H^*$ to the ground truth pose $H$ can then be expressed as the average absolute distance between individual body parts,

$$D(H, H^*) = \sum_{m=1}^{M} \frac{|h_m - h_m^*|}{M}. \tag{11}$$

For $N$ sequences of $T$ frames we can compute the average distance error using the following:

$$\xi = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} D\left(H_{t,n}, H_{t,n}^*\right). \tag{12}$$

Figure 12 shows the average error and computation time of 3D estimation results obtained by 40 iterations and 12-step burn-in period. We observe that the AG sampler performs better than the Gibbs sampler, and the two-stage approach performs better than classical one-stage approach, with lower distance error. It demonstrates that the annealed two-stage method is able to resolve the occlusion and hidden feature problem with increased accuracy. It seems that the two-stage approach requires more computational time, because more nodes need to be explored in the approach. However, the AG sampler takes less inference time for exploration. We



(a)

(b)

FIGURE 14: Average distance errors of the proposed two-stage method with respect to different burn-in numbers. (a) 3D distance errors, (b) 2D distance errors.

attribute this largely to that the sampler only explores the mode of the likelihood function, but not the entire likelihood space. Finally, the annealed two-stage method can achieve the best performance.

Algorithmic comparison was evaluated with the Bayesian framework of sequential importance resampling and annealed particle filtering, which is the baseline algorithm provided in the HumanEva for evaluation and comparison. An average distance error of 68 mm was obtained. Compared to the 51 mm average distance error of the proposed method, the annealed two-stage method can get better performance.

The behaviors of the two-stage method with respect to the iteration number and burn-in parameters were investigated. Figure 13 shows the effect of changing iteration parameter from 10 to 50. We observe that increased iteration number reduces distance errors. The decrease of distance errors converges at the 40 iterations. Different values of burn-in were also tested. Figure 14 shows the average distance error of 3D and 2D poses with respect to the burn-in periods with 40 iterations. In general, the performance of the proposed two-stage method is not sensitive to burn-in values. The errors for both Gibbs sampler and the AG sampler are fairly stabilized for 3D and 2D estimations. However, in all these burn-in parameters the estimation of AG sampler is consistently more accurate than that of Gibbs sampler.

## 6. Conclusions

We have presented a novel approach to solve the problem of 3D pose estimation from monocular image sequences. The approach proposes a two-stage inference hierarchy of Bayesian network. Visual features are extracted as the observations in the first stage of the network for the inference of 2D poses. The 2D poses along with assisted visual features are then applied for the inference of 3D poses. The EM algorithm is applied for the learning of network parameters. The advantages of the two-stage approach are twofold: First, it can improve the accuracy of 3D poses estimation; Second, both 2D and 3D poses are obtained by the same computational framework.

We introduce the use of simulated annealing as an effective mechanism to improve inferences. An annealed Gibbs sampler has been proposed for the inference of the two-stage Bayesian network. The sampler adopts an annealing process to explore the proposal distribution of sampling. The exploration can concentrate on the mode of the distribution and achieve more efficient computation. Experimental results demonstrate that the annealed two-stage approach has better accuracy and efficiency than one-stage method.

The approach currently has two limitations, first, the technique needs more accurate visual features to help the exploration of inferences. Second, temporal and spatial information are not incorporated into the approach to predict and smooth the inference result of each frame. Tracking algorithm and multiple cameras can further improve the estimation.

## References

[1] R. Poppe, "Vision-based human motion analysis: an overview," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[2] M. Brand, "Shadow puppetry," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '99)*, vol. 2, pp. 1237–1244, Kerkyyra, Greece, September 1999.

[3] A. Elgammal and C.-S. Lee, "Inferring 3d body pose from silhouettes using activity manifold learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. 681–688, Washington, DC, USA, June 2004.

[4] G. Loy, M. Eriksson, J. Sullivan, and S. Carlsson, "Monocular 3D reconstruction of human motion in long action

sequences," in *Proceedings of the European Conference on Computer Vision*, vol. 3024 of *Lecture Notes in Computer Science*, pp. 442–455, Prague, Czech Republic, 2004.

[5] R. Rosales and S. Sclaroff, "Inferring body pose without tracking body parts," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '00)*, vol. 2, pp. 721–727, Hilton Head Island, SC, USA, June 2000.

[6] E.-J. Ong and S. Gong, "A dynamic 3D human model using hybrid 2D-3D representations in hierarchical pca space," in *Proceedings of the British Machine Vision Conference*, pp. 33–42, Nottingham, UK, 1999.

[7] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," *Computer Vision and Image Understanding*, vol. 80, no. 3, pp. 349–363, 2000.

[8] A. Agarwal and B. Triggs, "A local basis representation for estimating human pose from cluttered images," in *Proceedings of the 7th Asian Conference on Computer Vision, Part I (ACCV '06)*, vol. 3851 of *Lecture Notes in Computer Science*, pp. 50–59, Hyderabad, India, January 2006.

[9] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 44–58, 2006.

[10] R. Bowden, T. A. Mitchell, and M. Sarhadi, "Non-linear statistical models for the 3D reconstruction of human pose and motion from monocular image sequences," *Image and Vision Computing*, vol. 18, no. 9, pp. 729–737, 2000.

[11] K. Rohr, "Towards model-based recognition of human movements in image sequences," *Computer Vision, Graphics and Image Proceeding: Image Understanding*, vol. 59, no. 1, pp. 94–115, 1994.

[12] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[13] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard, "Tracking loose-limbed people," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 421–428, Washington, DC, USA, June 2004.

[14] S.-F. Lan, M.-F. Ho, and C.-L. Huang, "Human motion parameter capturing using particle filter and nonparametric belief propagation," in *Proceedings of IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 37–40, Santa Fe, NM, USA, 2008.

[15] G. Hua, M.-H. Yang, and Y. Wu, "Learning to estimate human pose with data driven belief propagation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 747–754, San Diego, Calif, USA, June 2005.

[16] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Tracking people by their appearence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 65–80, 2007.

[17] M. W. Lee and I. Cohen, "A model-based approach for estimating human 3D poses in static images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 905–916, 2006.

[18] N. Thome, D. Merad, and S. Miguet, "Human body part labeling and tracking using graph matching theory," in *Proceedings of IEEE International Conference on Video and Signal Based Surveillance (AVSS '06)*, pp. 38–46, Sydney, Australia, November 2006.

[19] T. Leonid and T. Darrell, "Bayesian articulated tracking using single frame pose sampling," in *Proceedings of International Workshop on Statistical and Computational Theories of Vision*, Nice, France, 2003.

[20] N. Jojic, M. Turk, and T. S. Huang, "Tracking self-occluding articulated objects in dense display maps," in *Proceedings of International Conference on Computer Vision*, pp. 123–130, Kerkyra, Greece, 1999.

[21] N. Jojic and B. J. Frey, "Learning flexible sprites in video layers," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 199–206, Kauai, Hawaii, USA, December 2001.

[22] M. I. Jordan, *Learning in Graphical Models*, The MIT Press, Cambridge, Mass, USA, 1998.

[23] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[24] C. Borgelt and R. Kruse, *Graphical Models: Methods for Data Analysis and Mining*, Wiley, Chichester, UK, 2002.

[25] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[26] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[27] A. O. Bǎlan, L. Sigal, and M. J. Black, "A quantitative evaluation of video-based 3D person tracking," in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS '05)*, pp. 349–356, Beijing, China, October 2005.

[28] L. Sigal and M. Black, "HumanEva: synchronized video and motion capture dataset for evaluation of articulated human motion," Tech. Rep. CS-06-08, Brown University, 2006.

[29] L. M. Bergasa, M. Mazo, A. Gardel, M. A. Sotelo, and L. Boquete, "Unsupervised and adaptive Gaussian skin-color model," *Image and Vision Computing*, vol. 18, no. 12, pp. 987–1003, 2000.

[30] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.