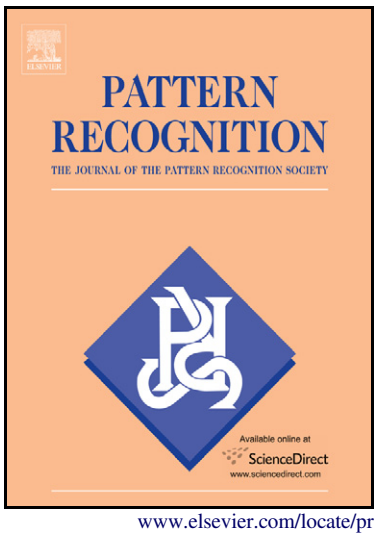


www.elsevier.com/locate/pr

Adaptive occlusion state estimation for human pose tracking under self-occlusions

Nam-Gyu Cho, Alan L. Yuille, Seong-Whan Lee

PII: S0031-3203(12)00398-6
DOI: http://dx.doi.org/10.1016/j.patcog.2012.09.006
Reference: PR4595

To appear in: *Pattern Recognition*

Received date: 27 September 2011
Revised date: 16 July 2012
Accepted date: 2 September 2012

Cite this article as: Nam-Gyu Cho, Alan L. Yuille, Seong-Whan Lee, Adaptive occlusion state estimation for human pose tracking under self-occlusions, *Pattern Recognition,* http://dx.doi.org/10.1016/j.patcog.2012.09.006

Adaptive Occlusion State Estimation for Human Pose Tracking under Self-Occlusions

Nam-Gyu Cho[a], Alan L. Yuille[a,b], Seong-Whan Lee[a,*]

[a]*Dept. of Brain and Cognitive Engineering, Korea University, Korea*
[b]*Dept. of Statistics, University of California, Los Angeles, CA 90095, USA*

Abstract

Tracking human poses in video can be considered as the process of inferring the positions of the body joints. Among various obstacles to this task, one of the most challenging is to deal with 'self-occlusion', where one body part occludes another one. In order to tackle this problem, a model must represent the self-occlusion between different body parts which leads to complex inference problems. In this paper, we propose a method which estimates occlusion states adaptively. A Markov random field is used to represent the occlusion relationship between human body parts in terms an occlusion state variable, which represents the depth order. To ensure efficient computation, inference is divided into two steps: a body pose inference step and an occlusion state inference step. We test our method using video sequences from the HumanEva dataset. We label the data to quantify how the relative depth ordering of parts, and hence the self-occlusion, changes during the video sequence. Then we demonstrate that our method can successfully track

*Corresponding Author: Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-713, Korea. Tel: (+82)-2-3290-3197, Fax: (+82)-2-3290-3583

Email addresses: ngcho@image.korea.ac.kr (Nam-Gyu Cho), yuille@stat.ucla.edu (Alan L. Yuille), swlee@image.korea.ac.kr (Seong-Whan Lee)

human poses even when there are frequent occlusion changes. We compare our approach to alternative methods including the state of the art approach which use multiple cameras.

Keywords: 3D human pose tracking, Computer vision, Self-occlusion

1. Introduction

The goal of human pose estimation is to find the human body configuration in 2D or 3D space from an input image. Pose tracking refers to the case when the input is an image sequence. Pose tracking has many potential applications such as human motion capture without using markers, Human Computer Interactions (HCI), Human Robot Interactions (HRI), video surveillance, etc. A marker-free human motion capture system has great advantages in an environmental setting, since it only uses one or a set of cameras, while a marker-based motion capture system requires not only a set of markers, but also special cameras (e.g., infrared cameras).

After pioneering work by Rohr [1] more than a hundred papers have been published on human pose estimation during the last two decades [2]. This research can be divided into two categories: discriminative approaches and generative approaches. Discriminative approaches learn mapping functions between the human pose and sets of features extracted from images. These methods can find the best matching pose quickly, and also give highly accurate results when the input image is similar to the training data. But, they have poor performance when the input image differs markedly from the training data. Model-based generative approaches use graphical models, e.g., Bayesian networks or Markov networks. In these approaches, graph nodes represent the state of a human body part and graph edges

model the relationships between the parts. Probability distributions are defined over this graphical structure to specify the probable poses of human figures and how they generate features in the image. Generative methods have the advantage that, unlike discriminative methods, they can deal with novel poses which the system has not been trained on. However, they suffer from exponentially increasing computational complexity of inference, largely due to the self-occlusion problem.

Self-occlusion means that one body part occludes another one, therefore, one body part will be overlapped by another one in the image. For example, it occurs when a human rotates with respect to the camera or a human is performing dynamic motion, such as boxing. Self-occlusion poses challenges for standard methods for object detection. For example, pictorial structures methods [3], which represent the human body as a set of linked rectangular regions, does not take self-occlusion into account. Sigal et al. [4] argue that the self-occlusion problem can be reduced by explicitly modeling it by an occlusion-sensitive likelihood model. This works well if the occlusion states (i.e. the depth ordering of parts) is known, for example if it is specified at the start of the motion and then does not change over time. But, in practice, the depth order of object parts – e.g., right arm, torso, and left arm – will change during a motion sequence, as we will quantify in our experiments.

In this paper, we propose an adaptive self-occlusion state estimation method that estimates not only the body configuration but also the occlusion states of body parts. The occlusion states are modeled as a state variable which takes three values and which represents the depth order between pairs of body parts, and hence enables adaptive estimation of the occlusion states. To simplify the combinatorial problem of estimating the occlusion states, we propose a novel inference scheme

that estimates the body pose and occlusion states separately. Our method is based on the following experimental observation: when the overlapping region (i.e. the self-occlusion) between body parts in the image is small, then pictorial structures [3] and the self-occlusion reasoning approach [4] give similar tracking performance. But as the overlapping region between two body parts expands, the tracking performance of [3] decreases while [4] maintains relatively high performance, provided the depth order is known and unchanging. From this observation, we postulate that if we can find overlapping body parts with small overlapping regions (where pictorial structures have fairly good performance) then we can estimate, and hence update, the occlusion state of overlapped body parts. This information can be used for the next inference step to simplify the search. In short, we use the occlusion estimates from the previous time frame to simply the estimation of body pose in the next frame, hence enabling us to re-estimate the occlusion states. This leads to efficient estimation of the occlusion state and prevents a possible combinatorial explosion.

2. Related Work

Estimating 2D human pose is difficult because of image noises (e.g., illumination variation and background clutter), self-occlusion, and the varieties of human appearances (e.g., clothing, gender, and body shape) [5, 6]. Estimating and tracking 3D human pose is even more challenging because of the large state space of the human body in 3D and our indirect knowledge of 3D depth [7].

Discriminative approaches proposed matching algorithms which specify mappings between 2D features extracted from the image and 3D object models. For example, Agarwal et al. [8] used shape context features which can be matched

to the 3D object models. Lee et al. [9] proposed a method that matches image descriptors to silhouettes of the 3D object. Raskin et al. [10] also matched to 3D silhouettes and proposed an annealed particle filtering method to perform the matching. But these approaches suffer from the ambiguities of the input images. Although the silhouette is fairly easy to extract, it has limited power to discriminate between possible 3D poses largely because of self-occlusion. Using multiple cameras can help handling self-occlusion problem by summing up all possible information from each camera [11, 12]. It also helps other fields such as action recognition [13, 14].

Model-based generative approaches use graphical models to represent humans and define a conditional probability distribution for the input image given the object pose. These graphical model approaches algorithms to infer the most probable 3D pose. Particle filters [15] is a standard way to perform inference using these models. In particular, Sminchisescu et al. [16] implemented a mixture density propagation approach for addressing the depth ambiguity problem from a single viewpoint. Bernier et al. [17] and Gupta et al. [18] used multiple cameras to overcome the depth ambiguity of input images, but these methods are inapplicable to video surveillance applications where there is typically only a single camera. Having a strong priors also can help for tracking motion [19]. However, it becomes hard when try to learn from various motions. In general, inference on graphical models suffer from computational complexity that increases exponentially as the number of nodes increases, although approximate methods like Belief Propagation (BP) can sometimes be used to reduce complexity [20]. Variants include non-parametric Belief Propagation (NBP), which represents distributions non-parametrically (e.g., avoiding Gaussian assumptions) and use Monte Carlo

sampling methods [17]. But typically large numbers of particles are required by this approach, as the complexity of the model increases, which becomes very computationally expensive although there are ways to improve this, such as Mean Shift Belief Propagation (MSBP) [21].

The use of generative models for objects for pose estimation and tracking is made challenging because of occlusion, which includes self-occlusion between different parts of the object. Occlusion not only complicate the modeling but also make the inference task considerably harder. It can result in inaccurate measurements of the object or the object parts. Recent methods suggest addressing the problem by using a state variable to represent the pseudo depth which depends on the object pose. This state variable enables us to improve the generative model by specifying which parts of the object are visible or invisible [22, 23].

For human pose estimation, the self-occlusion problem is the more challenging. Body parts tend to have similar visual appearance and so it becomes hard to distinguish them by image measurements (otherwise we could deal with self-occlusion by estimating which body parts are present directly from the image). Sigal et al. [4] developed an approach to human pose estimation building on previous work on hand tracking with self-occlusion [24]. Sigal et al. developed an 'occlusion-sensitive likelihood model', which used hidden variables to specify the visibility of pixels. Wang et al. [25] used this occlusion-sensitive likelihood model to estimate 2D human pose. They used a linear programming algorithm to get fast inference under self-occlusion. But all these methods assume the depth ordering of body parts is known in advance. Hence they cannot be directly applied to applications like tracking where the depth order of parts keeps changing (e.g., for boxing or dancing).

Our approach in this paper builds on Sigal et. al [4]. We use an explicit occlusion state variable that represents not only the depth order but also the visibility of different body parts. This enables us to relax the assumption that depth order, and part visibility, is unchanging. Instead we develop an adaptive inference algorithm which enables us to update the occlusion state variable while estimating the 3D body pose. This enables us to track human motion efficiently without making assumptions about the depth order.

3. The Adaptive Occlusion State Estimation Method

3.1. Overview of the Adaptive Occlusion State Estimation Method

The proposed adaptive occlusion state estimation method is formulated by a probabilistic graphical model with a corresponding 3D human model. Figure 1 gives an overview of the proposed method and Table 1 describes the terminology of the graphical model. In this paper, we propose a novel inference scheme which estimate the 3D body pose and the occlusion state Λ in alternation. The occlusion state variable estimated at time step $t-1$ is used to estimate the 3D body pose at time t which, in turn, is used to estimate the occlusion state at time t.

Estimating the occlusion state Λ is computationally challenging. In this paper we use 15 body parts to represent the human body and define 3 different occlusion states for describing the occlusion relationships between pairs of body parts, which yields 3^{14} ($\simeq 10^5$) possible occlusion states. These states affect the likelihood functions for the image observations, which are time consuming to compute. To address this challenge we exploit our experimental observation discussed in Section 1. We determine a 3D cylindrical human body model from our estimates of the body configuration and use this model to find those body parts which oc-

Table 1: The notations used in the proposed method.

Notation	Description
$X = \{X_1, ..., X_{15}\}$	set of nodes for body parts
$\mathbf{x}_i = (x, y, z)$	position of X_i in 3D space
$\boldsymbol{\theta}_i = (\theta_x, \theta_y, \theta_z)$	orientation of X_i in 3D space
$\Upsilon(X_i)$	set of pixels in the area in the image where X_i is projected
$W_i = \{w_i(u)\}, (u \in \Upsilon(X_i))$	set of visibility variables of pixel u's
$\Lambda = \{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_{15}\}$	set of occlusion state variables
$\boldsymbol{\lambda}_i = (\lambda_{i,1}, ...\lambda_{i,15}), \lambda_{i,i} = 0$	set of occlusion state variables between node X_i and the others
$E = (E_K, E_{O\|\Lambda}, E_T)$	set of edges
E_K	$X_i, X_j \in E_K$ such that $X_i, X_j \in X$
$E_{O\|\Lambda}$	$X_i, X_j \in E_{O\|\Lambda}$ such that $X_i, X_j \in X$
E_T	$X_i^{t-1}, X_i^t \in E_T$ such that $X_i^{t-1}, X_i^t \in (X^{t-1}, X^t)$
I	input image
$\nu_{i,j}$	indicator for overlapping body parts
ϕ_i	potential of observation
ψ_{ij}^K	potential of kinematic relationship
ψ_i^T	potential of temporal relationship

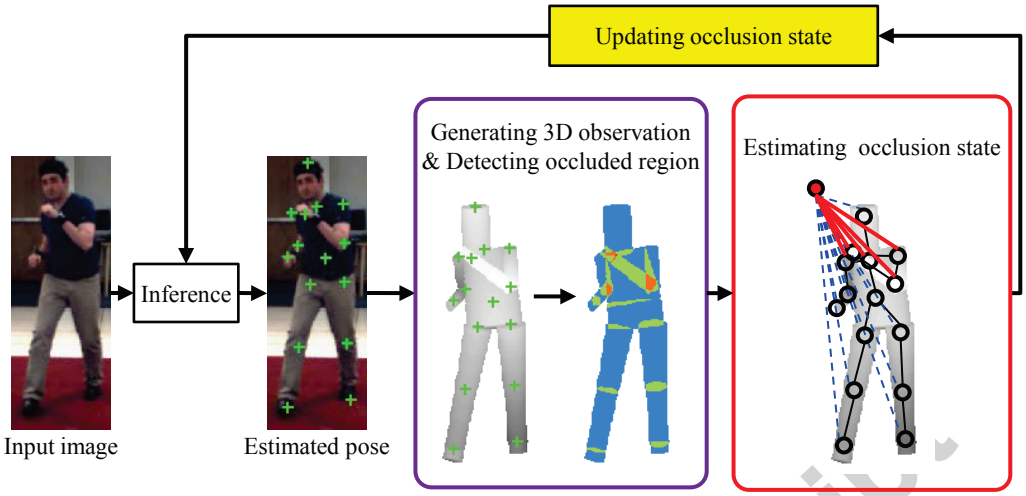


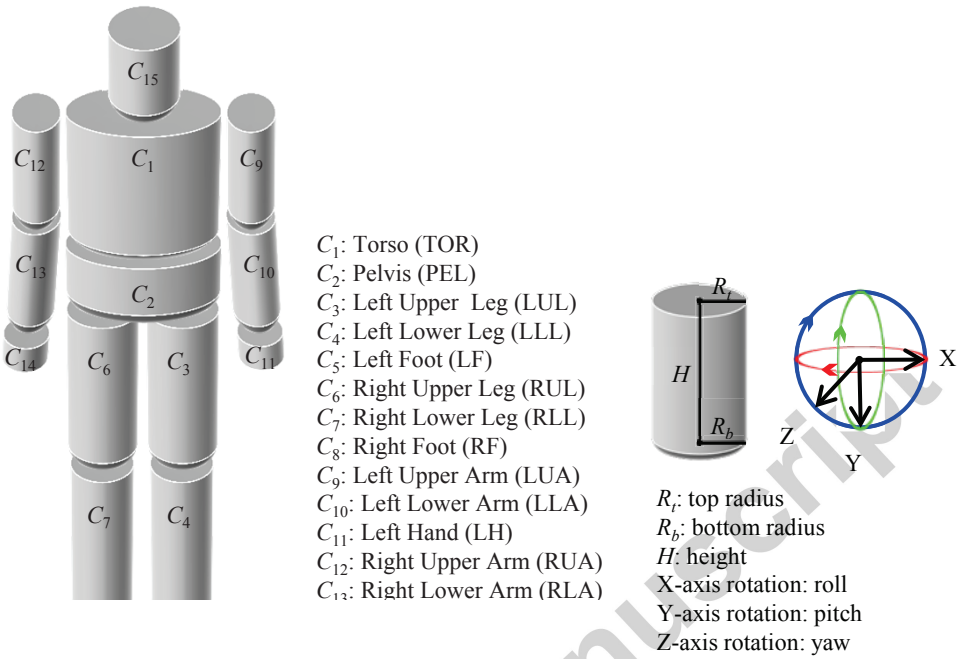
Figure 1: Overview of the adaptive occlusion state estimation method. There are two steps. At the first step of time t, the body pose $\hat{X}^t$ is estimated with respect to the occlusion states $\hat{\Lambda}^{t-1}$ that was estimated at the previous time $t-1$. At the second step of time t, overlapping body parts are detected from a 3D observation that was generated from $\hat{X}^t$. Then $\hat{\Lambda}^t$ is estimated from these detected body parts.

clude each other. The occlusion state is only estimated for these occluding body parts (see Section 3.3.4 and 3.3.5 for the details).

3.2. 3D Human Model

Figure 2 describes the 3D human model used in this paper. This 3D human model consists of 15 3D cylinders. Each cylinder has one of two types of DOFs (Degrees of Freedom): C_1, C_2, C_3, C_5, C_6, C_8, C_9, C_{11}, C_{12}, C_{14} and C_{15} have 3 DOF (rotation about the x, y, and z axes) and C_4, C_7, C_{10}, and C_{13} have 1 DOF (rotation about the x axis). Cylinder C_1 has 3 additional DOFs (the x, y, and z positions). The global position and orientation of the 3D human model is

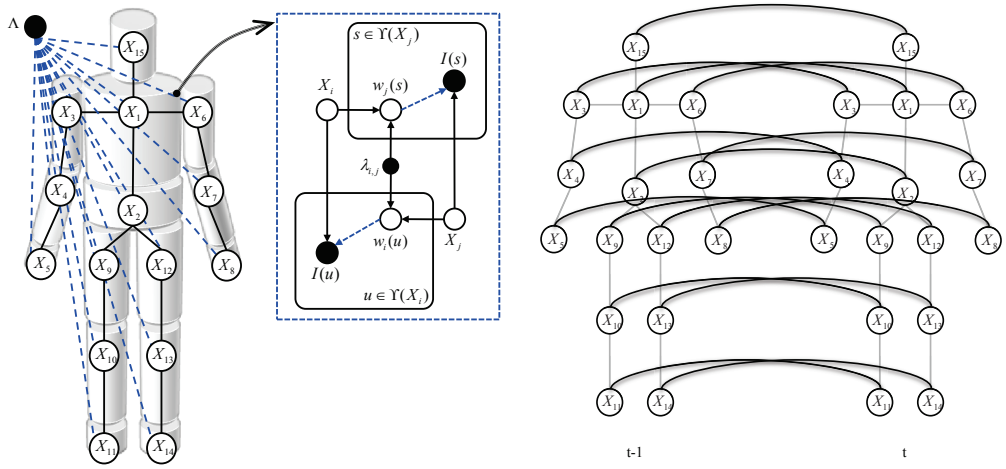
determined by the 6 DOFs of C_1.



(a) 3D human model. (b) 3D cylinder parameterization.

Figure 2: The 3D human model used in the proposed method. The names of the body parts and the parameterization of the 3D cylinders are described in 2(a) and 2(b) respectively.

3.3. MRF for Adaptive Occlusion State Estimation



(a) Graphical representation of the kinematic and occlusion relationship.

(b) The temporal relationship.

Figure 3: The MRF used for adaptive occlusion state estimation. The left panel of Figure 3(a): each node X_i ($i = 1, ..., 15$) corresponds to each C_i of the 3D human model in Figure 2(a). The solid lines represent the kinematic relationships between adjacent body parts and the dashed blue lines represent the occlusion relationships between body parts. The right panel of Figure 3(a): the graphical representation of the occlusion relationships encoded by blue edges with respect to the occlusion state variable $\lambda_{i,j}$. $w_i(u)$ represents the visibility of pixel u in $\Upsilon(X_i)$. $\Upsilon(X_i)$ is the area in the image where X_i is projected to. Figure 3(b): the temporal relationship is set between X_i^{t-1} and X_i^t.

We track the 3D human pose under self-occlusion using a Markov Random Field (MRF) with a state variable Λ representing the occlusion relationship between body parts. These are illustrated in Figure 3. The variable $w_i(u)$ is used to explain the visibility of pixel u, which is in the area of the image where X_i is projected. This variable depends on the occluders of X_i. Node X_i corresponds to C_i

of the 3D human model. The probability distribution over this graph is specified by the set of potentials defined over the set of edges. These edges are defined by the occlusion, kinematic, and temporal relationships between nodes. This probability distribution specifies the body configuration under self-occlusion.

The primary goal of visual tracking is to determine the posterior distribution $p(X^\tau|I^{1:\tau})$ for the current joint configuration of the model X^τ at the current time step τ, conditioned on all the input images $I^{1:\tau} = \{I^1, ..., I^\tau\}$ up to that time step [26]. In this paper, the posterior distribution of model X conditioned on all input images up to the current time step τ and occlusion state variable Λ is,

$$p(X^\tau|I^{1:\tau};\Lambda^{1:\tau}) = \frac{1}{Z}p(I^\tau|X^\tau;\Lambda^\tau)\int p(X^\tau|X^{\tau-1}) p(X^{\tau-1}|I^{1:\tau-1};\Lambda^{\tau-1})dX^{\tau-1}, \tag{1}$$

where Z is a normalization constant. This distribution is an example of a pairwise MRF defined over the time steps from 1 to τ [24],

$$p(X^\tau|I^{1:\tau};\Lambda^{1:\tau}) = \frac{1}{Z}\exp\left\{-\sum_{i\in X^{1:\tau}}\phi_i^C(I,X_i;\boldsymbol{\lambda}_i) - \sum_{ij\in E_K^{1:\tau}}\psi_{ij}^K(X_i,X_j) - \sum_{i\in E_T^{1:\tau},t\in 1:\tau}\psi_i^T(X_i^t,X_i^{t-1})\right\}, \tag{2}$$

where Z is a normalization constant. Details will be explained in following sections.

3.3.1. The Structure of MRF for the Proposed Method

The notations and descriptions of the MRF are listed in Table 1. Formally, the MRF is a graph $G = (V, E)$ where V is the set of nodes and E is the set of

edges. In this paper, the graph has state variables X, W and Λ. The edges are defined by the set of relationships (occlusion, kinematic, and temporal) and these relationships are modeled as the set of potentials. An occlusion relationship $E_{O|\Lambda}$ is formed between X_i and pixels u in $\Upsilon(X_i)$ with $w_i(u)$, and $\lambda_{i,j}$ where j is a possible occluder of X_i. The occlusion relationship changes its topology with respect to the occlusion state variable Λ, e.g., if node X_i and X_j has no occlusion relationship, the link between X_i and X_j (over related pixels and visibility variables) disappears. This will be explained in detail in the following section. The kinematic relationship E_K is defined over adjacent nodes. The temporal relationship E_T is formed over X_i^t and X_i^{t+1}.

3.3.2. The States of Nodes (Variables)

The state of X_i consists of the 3D position $\mathbf{x}_i$ and 3D orientation $\boldsymbol{\theta}_i$. $w_i(u)$ represents the visibility of pixel u generated by X_i and it has a binary state defined by,

$$w_i(u) = \begin{cases} 0, & \text{if pixel } u \text{ is occluded} \\ 1, & \text{if pixel } u \text{ is not occluded,} \end{cases} \tag{3}$$

where the value of this state variable is determined by the state of the occluders of X_i with respect to the state of $\boldsymbol{\lambda_i}$ (see Section 3.3.3). We introduce the state variable Λ to represent the occlusion relationship between body parts. As illustrated in Figure 4, $\lambda_{i,j}$ is only defined between different body parts (i.e. $\lambda_{i,i} = 0$). It represents one of three occlusion states between two different body parts X_i and X_j. State 1 indicates that neither of the two body parts occludes the other one, state 2 indicates that body part X_i occludes body part X_j, and state 3 indicates that body part X_i is occluded by body part X_j. The topology of G (actually $E_{O|\Lambda}$) is changed with respect to the state of $\lambda_{i,j}$. When $\lambda_{i,j} = 1$, sets of pixels defined

by the states of X_i and X_j are independent, and observation potentials of X_i can be calculated without considering occluder X_j's. Therefore, the edge between X_i and $w_j(s)$, and the edge between X_j and $w_i(u)$ disappear. When $\lambda_{i,j} = 2$, only those pixels of X_i that lie in the overlapping area are dependent on the pixels of X_j (i.e., the edge between X_i and $w_j(s)$ disappears). When $\lambda_{i,j} = 3$, the dependency is changed inversely. Consequently, in terms of adaptivity, the proposed occlusion state adapts to the changes of self-occlusions in the input image by changing its topology.

3.3.3. The Potentials

In order to define a conditional probability distribution for the human body configuration X given the input image I, we use the three potentials listed below in Table 1 in Section 3.3.1.

Observation Potential. The observation potential is calculated with respect to the occlusion relationship between body parts. That is, a body part located at the top of the depth order, calculated from Λ, is calculated first. We use two image cues, color and edges, to calculate the observation potential as follows:

$$\phi_i(I, X_i; \boldsymbol{\lambda}_i) = \phi_i^C(I, X_i; \boldsymbol{\lambda}_i) + \phi_i^E(I, X_i; \boldsymbol{\lambda}_i), \tag{4}$$

where ϕ_i^C is the observation potential for the color cue and ϕ_i^E is the observation potential for the edge cue. We modified the occlusion-sensitive likelihood model [4] for the observation potential with respect to the occlusion state variable Λ. An important issue of the occlusion-sensitive likelihood model is how to find the configuration of W_i, the set of visibility variables about X_i, in order to measure the observation potential. The depth order for the current image was assumed to be known in previous studies [4] [24] [25]. However, using the proposed occlusion

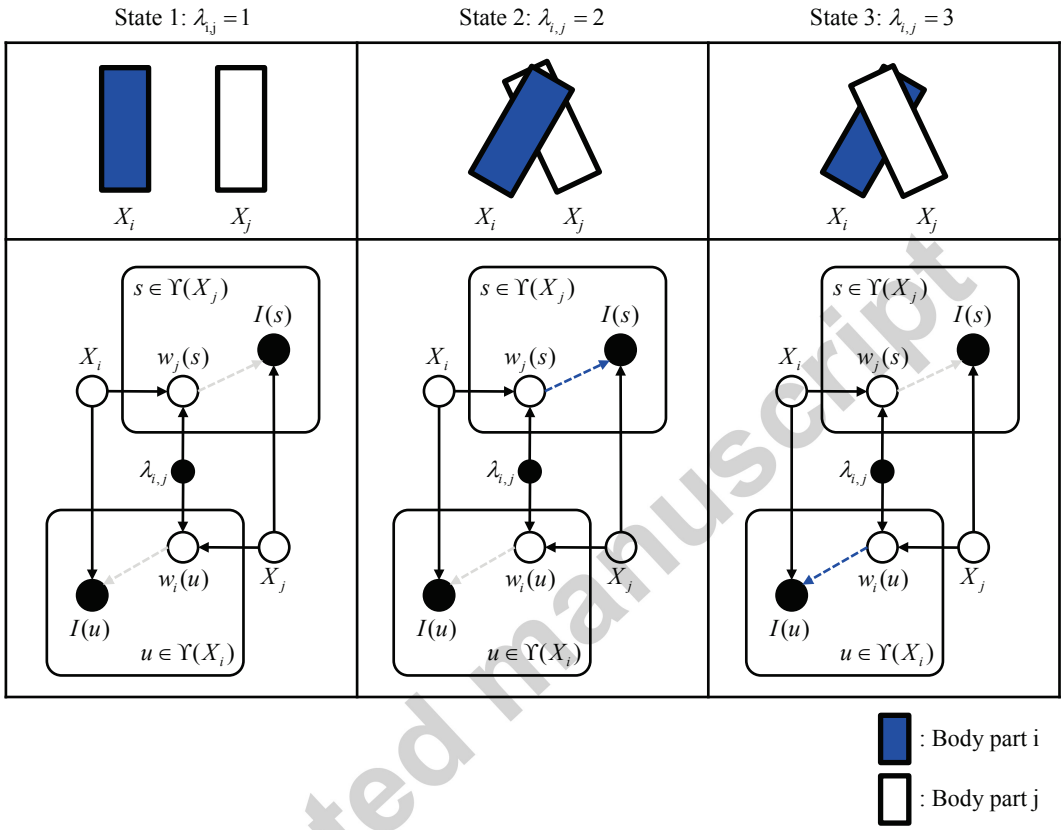


Figure 4: Definition of three occlusion states between two body parts. The topology of $E_{O|\Lambda}$ (dashed edges in blue) changes as the value of $\lambda_{i,j}$ changes. For example, when $\lambda_{i,j} = 2$ information (probability) of node X_j about pixel u is not necessary for calculating the probability of node X_i about pixel u. This relationship is represented by the light gray dashed edge.

Figure 5: An example of self-occlusion between body part j and i (j = TOR and i = RLA) where i is occluded by j. The green and gray regions are non-overlapping regions of X_j and X_i respectively. The yellow region is the overlapping region of X_j and X_i.

state variable Λ, the configuration of W_i can be calculated deterministically. For example, in Figure 5, X_i is occluded by X_j (i is RLA and j is TOR). In this case, we assume that ϕ_j is calculated in advance. The configuration of W_i is determined by the calculated overlapping region of X_i and X_j. Therefore, W_i can be represented as,

$$W_i = \{w_i(u'), w_i(u)\}, \tag{5}$$

where $w_i(u') = 0$ for $u' \in \Upsilon'(X_i)$ where $\Upsilon'(X_i) = (\Upsilon(X_j) \cap \Upsilon(X_i))$. And $w_i(u) = 1$ for $u \in (\Upsilon(X_i) - \Upsilon'(X_i))$. This leads to separate calculations of the observation potential. The observation potential for the color cue is formulated as follows:

$$\phi_i^C(I, X_i; \boldsymbol{\lambda}_i) = \phi_i^{C_{visible}}(I, X_i; \boldsymbol{\lambda}_i) + \phi_i^{C_{occluded}}(I, X_i; \boldsymbol{\lambda}_i), \tag{6}$$

where the first term is for the visible area, and the second term is for the occluded

area. The visible term is formulated as,

$$\phi_i^{C_{visible}}(I, X_i; \boldsymbol{\lambda}_i) = \prod_{u \in (\Upsilon(X_i) - \Upsilon'(X_i))} p_C(I_u), \tag{7}$$

where $\Upsilon'(X_i) = (\Upsilon(X_i) \cap \Upsilon(X_j))$ and the pixel probability is,

$$p_C(I_u) = \frac{p(I_u|\text{foreground})}{p(I_u|\text{background})}, \tag{8}$$

where $p(I_u|\text{foreground})$ and $p(I_u|\text{background})$ are the distributions of the color of pixel u given the foreground and background. These distributions are learned from the foreground and background image patches of the data set. The occluded term is formulated as,

$$\phi_i^{C_{occluded}}(I, X_i; \boldsymbol{\lambda}_i) = \prod_{u' \in \Upsilon'(X_i)} [z_i(I_{u'}) + (1 - z_i(I_{u'}))p_C(I_{u'})], \tag{9}$$

and $z_i(I_{u'})$ is calculated as follows,

$$z_i(I_{u'}) = \frac{1}{N_O} \sum_{\forall X_j \text{ s.t. } \lambda_{i,j}=4} \phi_j^C(I(u'), X_j^s; \boldsymbol{\lambda}_i), \tag{10}$$

where N_O is the total number of parts that occlude part i.

Kinematic Potential. We model the kinematic relationship between two adjacent parts using kinesiology, also known as human kinetics, which defines the Range of Motion (ROM) of human joints [27]. Table 2 shows an example of ROM for arm joints. In this paper, we use ROM to approximate the possible range of orientation of adjacent body parts in 3D space, e.g., the ROM of shoulder flexion-hyperextension is converted to the rotation range [-50 180] on the x-axis and abduction-adduction is converted to the rotation range [-50 180] on the z-axis. ROM is relative to its parent joint. In other words, since the human body can be represented as a kinematic tree, the orientation of a joint is determined relative to

Table 2: The example of ROM of joints related to arm

Joint/Segment	Movement	Range (degree)
Shoulder	Flexion	180
	Hyperextension	50
	Abduction	180
	Adduction	50
Elbow	Flexion	140
	Hyperextension	0
Forearm	Pronation	80
	Supination	80
Wrist	Extension (dorsiflexion)	60
	Flexion (palmar flexion)	60

its parent joint. This kinematic relationship of the position and orientation of two adjacent body parts is formulated as,

$$\psi_{ij}^{K}(X_i, X_j) = N(d(\mathbf{x}_i, \mathbf{x}_j); \mu_k, \sigma_K) f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j), \tag{11}$$

where X_i, $X_j \in E_K$ $(i < j)$ and $d(\mathbf{x}_i, \mathbf{x}_j)$ is the Euclidean distance between X_i's proximal joint and X_j's distal joint (in this case, $i < j$). $N()$ is the normal distribution with $\mu_k (= 0)$ and standard deviation σ_K to allow adjacent body parts to be loosely linked and $f()$ is,

$$f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \begin{cases} 1, \text{if } LowerBound_{ij} \leq \boldsymbol{\theta}_i - \boldsymbol{\theta}_j \leq UpperBound_{ij} \\ 0, \text{otherwise} \end{cases}, \tag{12}$$

where $LowerBound_{ij}$ and $UpperBound_{ij}$ are the lower and upper bound of ROM between X_i and X_j defined by kinesiology. For more accurate tracking, we learn

the distribution of kinesiology from the HumanEva data set et al. [28].

Temporal Potential. This potential models the temporal relationship of a part between two consecutive time steps $t-1$ and t as a Gaussian distribution,

$$\psi_i^T(X_i^t, X_i^{t-1}) = p(X_i^t - X_i^{t-1}; \boldsymbol{\mu}_i, \Sigma_i), \tag{13}$$

where $\boldsymbol{\mu}_i$ is the dynamics of X_i at the previous time step and Σ_i is a diagonal matrix with a diagonal elementis identical to $|\boldsymbol{\mu}_i|$. Both the kinematic and temporal edges are illustrated in Figure 3.

3.3.4. Inference: Body Configuration Estimation

The goal of inference is to obtain the best state of X and Λ from the conditional probability distribution in (2) at time step t. In this paper, we use two steps to estimate the 3D body configuration and occlusion state variable separately. We can assume that Λ was estimated at the previous time step $t-1$ as $\hat{\Lambda}^{t-1}$ and this is given in order to estimate the body configuration $\hat{X}^t$ from the input image at the current time step t. This is formulated as follows,

$$\hat{X}^t = \arg\max_{X^t} p(X^t | I^{1:t}; \hat{\Lambda}^{t-1}). \tag{14}$$

In order to perform efficient inference, we use the Belief Propagation (BP) algorithm. BP uses local messages that sum up the entire set of probabilities about neighbor nodes with regard to their states. Two types of messages are used: kinematic and temporal. The kinematic message, which propagates from X_i^t to X_j^t about the kinematic relationship, is calculated as,

$$\begin{aligned} m_{ij}^K(X_j^t) = \sum_{X_i^t} \exp\left\{\phi(I^t, X_i^t; \hat{\boldsymbol{\lambda}}_i^{t-1}) + \psi_{ij}^K(X_i^t, X_j^t)\right\} \\ m_i^T(X_i^t) \prod_{k \in N(i) \setminus j} m_{ki}^K(X_i^t), \end{aligned} \tag{15}$$

where $N(i)\backslash j$ is neighbors of i except j. The temporal message, which propagates information from X_i^{t-1} to the X_i^t about the temporal relationship, is calculated as,

$$m_i^T(X_i^t) = \sum_{X_i^{t-1}} \exp\left\{\phi(I^{t-1}, X_i^{t-1}; \hat{\boldsymbol{\lambda}}_i^{t-2}) + \psi_i^T(X_i^t, X_i^{t-1})\right\} \\ m_i^T(X_i^{t-1}) \prod_{k \in N(i)} m_{ki}^K(X_i^{t-1}) \quad , \tag{16}$$

where $m_i^T(X_i^{t-1})$ is the temporal message from X_i^{t-2} to X_i^{t-1}. The belief of X_i^t is formulated as,

$$b_i(X_i^t) = \exp\left\{\phi_i(I^t, X_i^t; \hat{\boldsymbol{\lambda}}^{t-1})\right\} m_i^T(X_i^t) \prod_{k \in N(i)} m_{ki}^K(X_i^t), \tag{17}$$

and this belief approximates the marginal distribution.

We use the SIR (Sequential Importance Resampling) principle to represent the posterior distribution (which approximately equals the belief) of each body part by a set of N random samples with corresponding weights. The weights of the samples are normalized and propagated over time using the temporal model and then newly assigned with respect to their likelihood function [28]. In order to find the modes of the posterior distribution better, we iterate the SIR steps twice to infer the body configuration at each time step.

Tracking human poses using graphical models such as MRFs requires using BP over the whole time sequence. This requires performing inference on a graph defined the entire time sequence, but this is computationally too expensive. Therefore, we conduct BP over two consecutive time steps $t-1$ and t.

3.3.5. Occlusion State Variable Estimation

In this section, we describe how the occlusion state variable is estimated as one of three states. First we have to consider the combination of occlusion states

among the 15 body parts. There are 3^{14} ($\simeq 10^5$) possible combinations of occlusion state variable Λ^t and this increases the complexity further (recall Section 3.1). In this paper, we propose a novel occlusion state estimation method based on our experimental observation (see Section 1). In order to find overlapping (occluding) body parts, we first define a criterion, similar to [29], for detecting overlapping body parts as follows,

$$\nu_{i,j} = \begin{cases} 1, \text{if } max\left(\dfrac{\Upsilon(\hat{X}_i^t) \cap \Upsilon(\hat{X}_j^t)}{\Upsilon(\hat{X}_i^t)}, \dfrac{\Upsilon(\hat{X}_i^t) \cap \Upsilon(\hat{X}_j^t)}{\Upsilon(\hat{X}_j^t)}\right) \geq T_0 \\ 0, \text{otherwise} \end{cases}, \tag{18}$$

where $\Upsilon(\hat{X}_i^t)$ is the set of pixels in the area of the image where the estimate of X_i^t is projected. T_0 is a threshold determined empirically as 0.15. $\nu_{i,j}$ is an indicator for occluding body parts. If the value of $\nu_{i,j}$ is set to 0, the value of $\lambda_{i,j}^t$ is set to 1. Otherwise, $\lambda_{i,j}^t$ is estimated using the following equation,

$$\hat{\lambda}_{i,j}^t = \underset{\lambda_{i,j} \in \{2,3\}}{\arg\max}\, \phi(I^t, \hat{X}_i^t; \lambda_{i,j}), \tag{19}$$

where $\hat{X}_i^t$ is the estimate of X_i^t from the previous step. This process is illustrated in Figure 6.

3.3.6. Proposals

In human body pose estimation, strong priors improve the performance robustness, but also have limitations [30]. Robust body part detectors, e.g., for head, torso, and limbs, ensure that the pose estimation task is easier. This reduces the search space, but it is not always reliable, which is mainly due to the image noise and self-occlusions [31, 32]. In this paper, we construct proposals for the head and torso: a face detector [33] and a head-shoulder contour detector for the torso

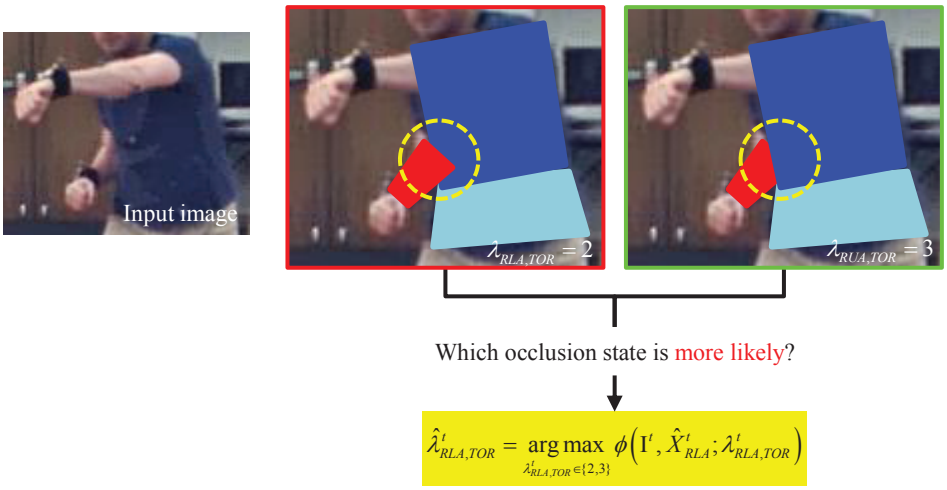


Figure 6: Process of occlusion state estimation. In order to estimate the occlusion state $\lambda^t_{RLA,TOR}$ as either of 2 or 3, $\phi(I^t, X^t_{RLA}; \lambda^t_{RLA,TOR})$ is calculated twice with respect to two the different values of $\lambda^t_{RLA,TOR}$. Then the state that has maximum value is assigned to the value of $\lambda^t_{RLA,TOR}$.

[31]. 50 samples of each part that have the most likely states are selected for the proposals, and these proposals are provided to the first step of body configuration inference (see Section 3.3.4).

4. Experimental Results and Analysis

The HumanEva-I data set [28] is used for evaluation. It contains 6 different motions: Walking, Jogging, ThrowCatch, Gestures, Boxing, and Combo. Each motion is performed by 4 subjects and recorded by 7 cameras (3 RGB and 4 gray scale cameras) with the ground truth data of human joints. We evaluate the performance on the first 5 test motions from HumanEva-I of subject S2 and camera C1. In order to evaluate the performance of occlusion state estimation, we hand-labeled the ground truth of the occlusion states for test motions. On average, manually specifying the occlusion states takes three minutes per image. Figure 7 shows how the ground truth of occlusion state is specified.

We compare the proposed method, Adaptive Occlusion State Estimation (AOSE), with the four state-of-the-art methods in the literature: Pictorial Structure (PS) [3], Self-occlusion Reasoning (SR) [4], Mixture of Factor Analyzers (FMA) [12], and Gait Generative Model (GGM) [34]. PS and SR take different approaches to tackle self-occlusion. Unlike SR and AOSE, PS does not explicitly model self-occlusion. Meanwhile, SR includes a special node for self-occlusion in the model, but it is assumed that the depth order with regard to self-occlusion is fixed and known a priori. The proposed AOSE alleviates this limitation by adaptively estimating the occlusion state in a frame. FMA tracks 3D human pose on a manifold space using multi-view information (camera C1-C3) while PS, SR, GGM, and AOSE use a single view (camera C1). GGM tracks 3D human walking motion using two

	TOR	PEL	LUL	LLL	LF	RUL	RLL	RF	LUA	LLA	LH	RUA	RLA	RH	HED
TOR	0	1	1	1	1	1	1	1	4	4	1	1	1	1	1
PEL	1	0	1	1	1	1	1	1	1	4	4	1	1	3	1
LUL	1	1	0	1	1	1	1	1	1	1	4	1	1	1	1
LLL	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
LF	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
RUL	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
RLL	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
RF	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
LUA	3	1	1	1	1	1	1	1	0	1	1	1	1	1	1
LLA	3	3	1	1	1	1	1	1	1	0	1	1	1	1	1
LH	1	3	3	1	1	1	1	1	1	1	0	1	1	1	1
RUA	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
RLA	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
RH	1	4	1	1	1	1	1	1	1	1	1	1	1	0	1
HED	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Figure 7: Illustration of how occlusion states are specified. In the upper right matrix image, occlusion part pairs, e.g., TOR-LUA and LUA-TOR have occlusion state value 4 (red cell for occluded one) and 3 (green cell for occluder) respectively. In this manner, every part pairs get corresponding occlusion state values. Abbreviations for body parts are: TOR - torso, PEL - pelvis, LUL - left upper leg, LLL - left lower leg, LF - left foot, RUL - right upper leg, RLL - right lower leg, RF - right foot, LUA - left upper arm, LLA - left lower arm, LH - left hand, RUA - right upper arm, RLA - right lower arm, RH - right hand, and HED - head.

kinematic and visual manifolds.

PS, SR, and AOSE were implemented in MATLAB R2009a on a desktop PC (Intel core 2 Quad 2.66GHz CPU, 4GB RAM, and 64bit Windows 7 operating system). During the experiment, initializations were done manually for PS, SR, and AOSE. For body configuration inference, as described in Section 3.3.4, two SIR iterations are conducted with 50 samples and BP is conducted over two consecutive time steps using the UGM (Undirected Graphical Model) toolbox [35].

Table 3: The mean (and standard deviation) of tracking errors in millimeters over 150 frames of each motion.

Motion	PS [3]	SR [4]	FMA [12]	GGM [34]	AOSE
Walking	230.66 (209.85)	144.74 (76.05)	68.67 (24.66)	99.04	88.03 (42.71)
Jogging	113.99 (73.40)	164.02 (144.11)	72.14 (54.66)	-	53.58 (17.95)
ThrowCatch	164.83 (122.25)	96.77 (47.62)	68.03 (22.18)	-	99.30 (36.80)
Gestures	92.27 (34.97)	125.18 (69.83)	67.66 (23.85)	-	83.76 (36.05)
Boxing	208.44 (73.78)	152.60 (86.20)	70.02 (22.74)	-	89.48 (27.29)
Mean	155.39 (100.85)	120.97 (69.02)	69.30 (29.62)	99.04	82.83 (32.16)

4.1. Pose Tracking

In Table 3, the mean and the standard deviation of tracker error of the five tracking algorithms are reported. The error is measured as the absolute Euclidean distance in millimeters between the ground truth and estimated fifteen 3D joints (marker) positions on the body parts as reported in FMA [12]. Since it is impossible to estimate an invisible part from a single view image, we do not count the error from completely occluded parts (for PS, SR, and AOSE). On an average, taking these invisible part into the calculation gives roughly 10mm higher mean error.

Overall, the proposed method outperforms PS and SR for the entire motions except ThrowCatch. Because FMA tracks human pose using multi-views (3 cameras), it can exploit more information such as appearance cues under self-occlusion (it is possible to argue that there is no self-occlusion for multi-views since a part occluded in one camera can be seen by the other cameras). But PS, SR, and AOSE track from a single view. So it is not surprising that FMA shows a slightly better performance.

Table 4: Limb definition.

Each limb	Components
L-ARM	LUA (left upper arm), LLA (left lower arm), and LH (left hand)
R-ARM	RUA (right upper arm), RLA (right lower arm), and RH (right hand)
L-LEG	LUL (left upper leg), LLL (left lower leg), and LF (left foot)
R-LEG	RUL (right upper leg), RLL (right lower leg), and RF (right foot)

As shown in Figure 9, PS, SR, and AOSE show similar performance for the first few frames. But then performance of PS and SR starts degrading when the human pose alters significantly causing depth order changes. For example, in Figure 8, at frame 60 of Boxing, the depth order between the right arm and the torso has changed compared to frame 30. The tracking error for individual parts are plotted in Figure 10. In general, the tracking error of three methods for limb extremities – feet and hands – is higher than for other parts.

4.2. Occlusion State Estimation

In Table 5, we analyze the complexity of 5 test motions in terms of the number of occlusion state change (blue colored cell). The analysis is done for: (1) the whole body (averaging over all parts) to give a global complexity measure, and (2) for 4 limbs to measure the local complexity (see the definition of limbs in Table 4). We use the mean frame interval per occlusion state change (FOC) as a complexity measure. According to this measure, both globally and locally, Jogging is the most complex motion in the dataset and, on average, the occlusion state of the whole body changes every 1.84 frames. Note that the mean FOC for L-LEG and R-LEG in the ThrowCatch, Gestures, and Boxing motions cannot be calculated since subject S2 performs the motions in a standing pose (there is no occlusion

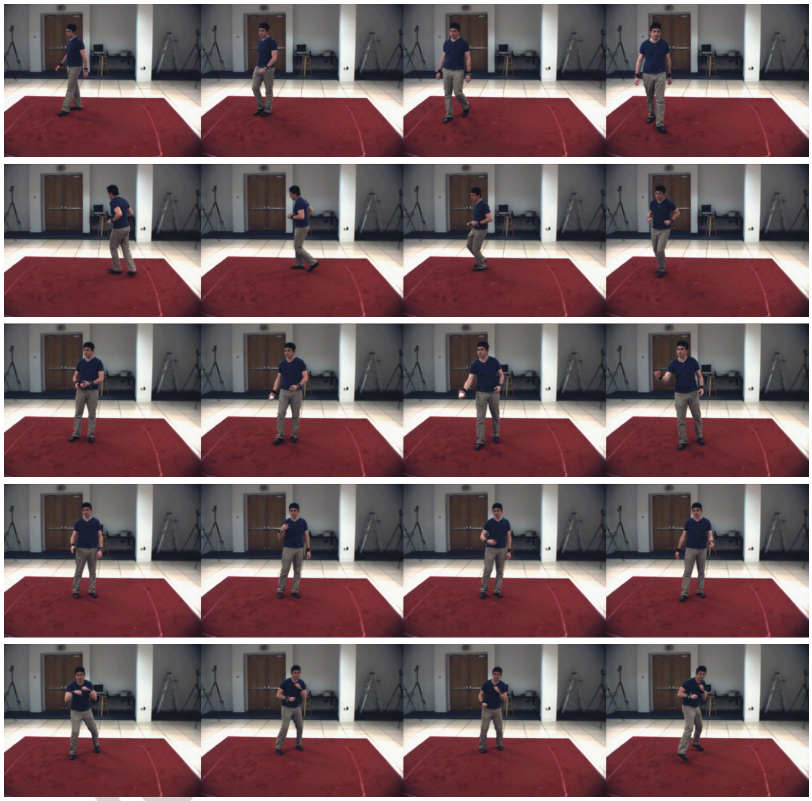


Figure 8: Image frame 30, 60, 90, 120, and 150 of (from top to bottom row) Walking, Jogging, ThrowCatch, Gestures, and Boxing motions. Jogging is the most complex motions in HumanEva-I dataset. Boxing is the second complex motion and Gestures is the least complex motion (see Table 5 in Section 4.2 for detailed information).

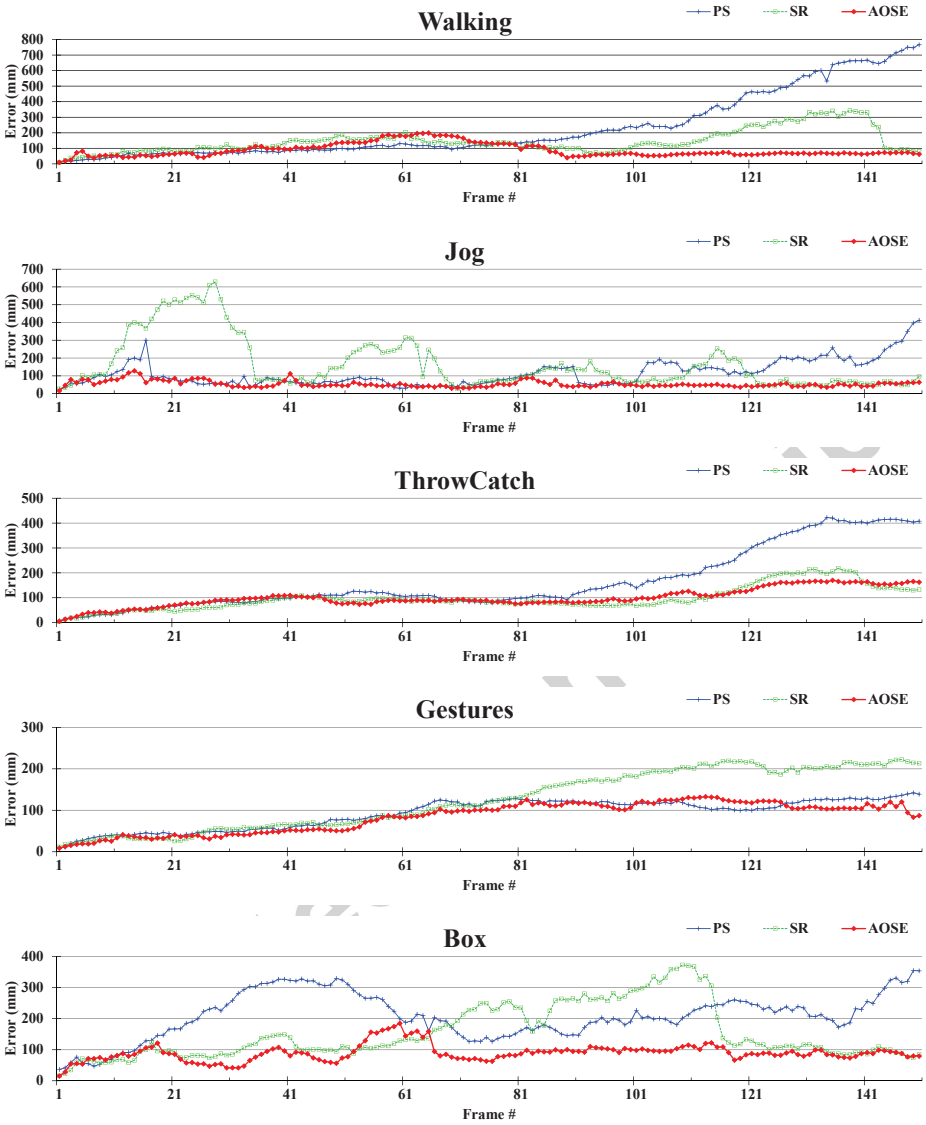


Figure 9: Tracking performance for 5 motions of the HumanEva-I dataset. Generally, PS loses body tracking under self-occlusion, and SR does after depth order changes (see Figure 8 while AOSE shows stable tracking performance with adaptively estimating occlusion state.

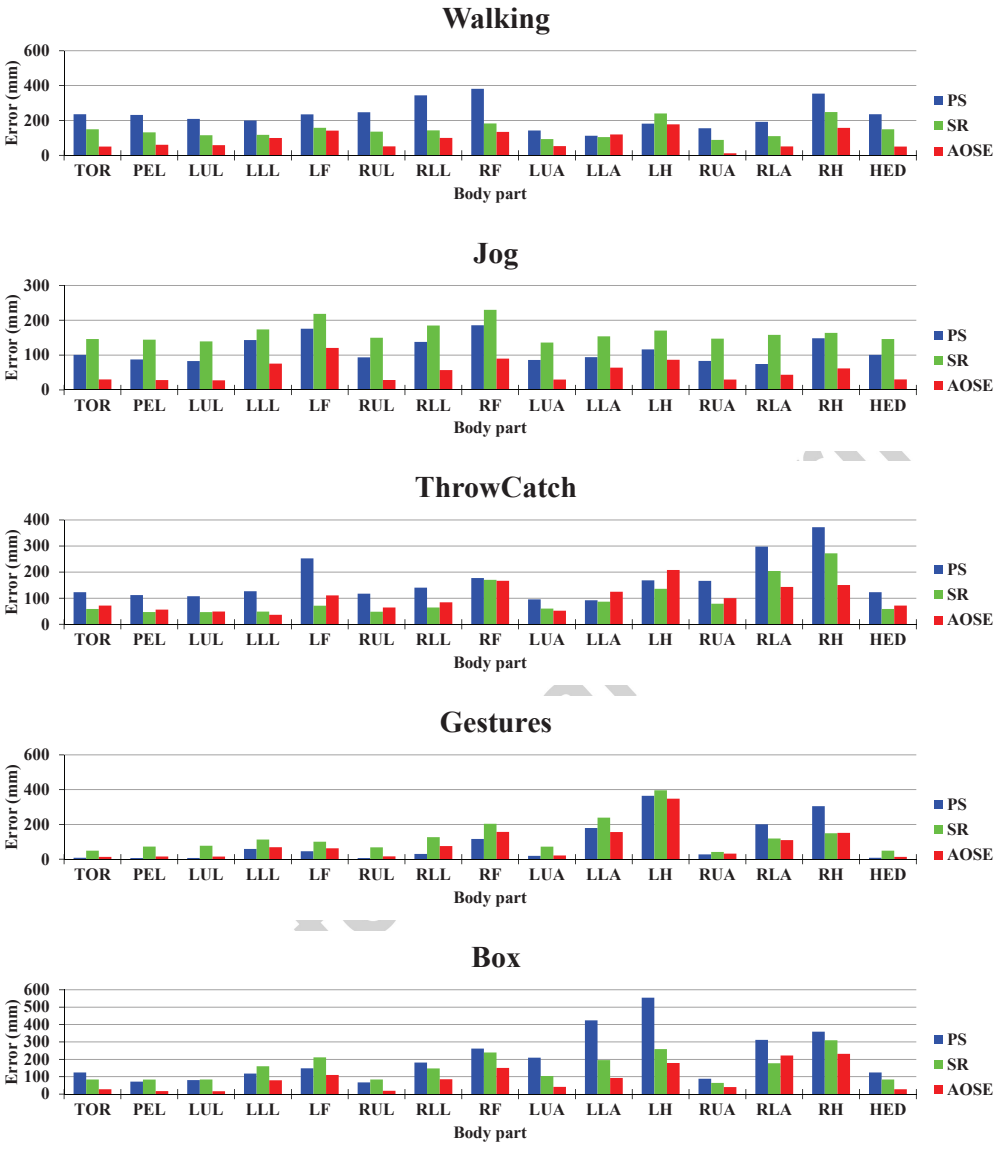
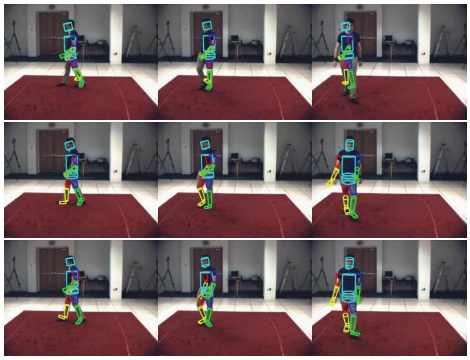
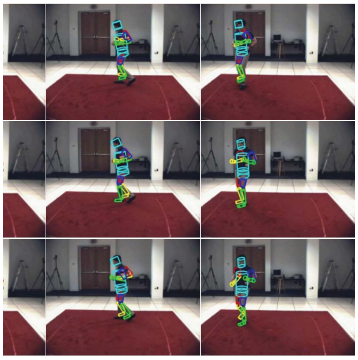


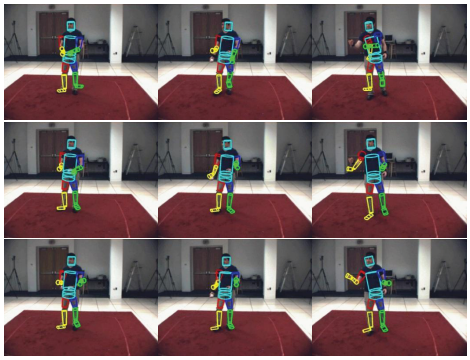
Figure 10: Tracking performance of individual parts. AOSE shows lower tracking error among three methods. By and large, three methods show relatively big error for both hands and feet.



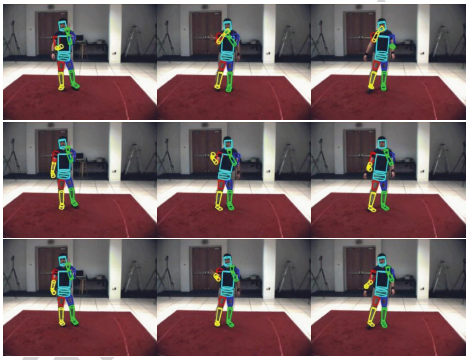
(a) Walking.



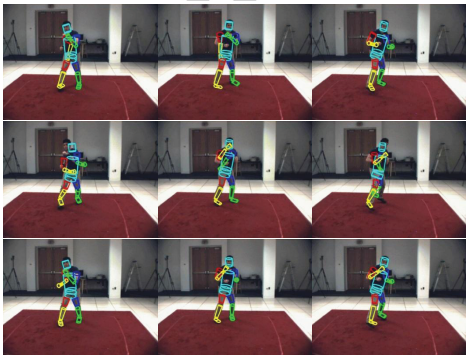
(b) Jogging.



(c) ThrowCatch.



(d) Gestures.



(e) Boxing.

Figure 11: At each subfigure, sample tracking result from three methods and three frames – PS (1st row), SR (2nd row), and AOSE (3rd row), and frame 30 (1st column), 90 (2nd column), and 150 (3rd column) from test video sequence of S2 – are shown.

Table 5: Motion complexity analysis of HumanEva-I (subject S2 and camera C1) and performance of occlusion state estimation.

Motion	Criterion	Whole body	Limb			
			L-ARM	R-ARM	L-LEG	R-LEG
Walking	Mean FOC	3.40	5.71	6.35	6.46	6.31
	Mean error (%)	9.13	15.08	10.26	8.84	10.56
Jog	Mean FOC	1.84	2.93	3.18	5.73	5.96
	Mean error (%)	4.25	17.38	17.38	8.07	11.44
ThrowCatch	Mean FOC	5.54	4.31	16.63	-	-
	Mean error (%)	3.34	23.68	17.66	-	-
Gestures	Mean FOC	8.60	43.00	10.33	-	-
	Mean error (%)	2.65	15.38	17.16	-	-
Box	Mean FOC	2.80	4.45	3.32	-	-
	Mean error (%)	2.83	25.48	22.17	-	-

state change for these parts).

The mean error of the occlusion estimation is calculated as follows,

$$E^t_{ose} = \frac{\sum_{i=1}^{K} Diff(\Lambda^t_i, \hat{\Lambda}^t_i)}{K} \tag{20}$$

where Λ^t_i is i^{th} element of the ground truth data of the occlusion state at time step t and $\hat{\Lambda}^t_i$ is i^{th} element of the estimate of the occlusion state at time step t. In eqn. (20) only the upper triangular part of Λ (and $\hat{\Lambda}$) is considered because Λ is symmetric (in the sense that $\lambda_{i,j}$ and $\lambda_{j,i}$ yield the same information). $Diff(a, b)$ returns 0 if a and b have the same value, otherwise, it returns 1. K is the total number of elements in the upper triangular matrix of Λ. $K = n \times (L - 1) - n$ where L is the total number of body parts and n is the number of body parts of limb.

As can be seen in Table 5, the proposed method shows good occlusion state estimation performance for both whole body and four limbs. Based on this result, we can say that our method has advantages not only for tracking but also for estimating the occlusion states of complex motion such as Walking and Jogging. In particular, it shows even better tracking performance then FMA [12] which uses multi-view informations for Jogging motion (see Table 3). However, it does not show good performance for L-ARM and R-ARM of the ThrowCatch and Boxing motions which are not complex in terms of FOC. Because the ThrowCatch and Boxing motions contain a pose which makes one of body parts be oriented parallel to the camera view – e.g., throw a punch towards to the camera – our current appearance model cannot represent this accurately.

5. Conclusions and Future Work

In this paper, we proposed an adaptive occlusion state estimation method for 3D human body tracking. Our method successfully tracks without assuming a known and fixed depth order. The proposed method can infer state variables efficiently because it separates the estimation procedure into body configuration estimation and occlusion state estimation. More specifically, in the occlusion state estimation step, we first detect body parts having an occlusion relationship using the overlapping body parts detection criterion. Then we estimated the occlusion states only for these overlapping body parts. This leads to an efficient state estimation algorithm. Experimental results carried on 5 motions of HumanEva-I dataset showed that the proposed method successfully tracks the 3D human pose and estimates the occlusion states in the presence of self-occlusion. The proposed method outperforms three competing methods for Jogging and has the second best

performance for the remaining test motions except ThrowCatch. We quantified the occlusion complexity of the motions sequences using the FOC measure and used this to determine how successful our method was for estimating occlusions.

We conjecture that the proposed method can be extended for tracking poses from (two or more) interacting people. Tracking poses of interacting people, however, will involve more complex problems such as dealing with more variable motion, inter-person occlusions, and possible appearance similarity of different people.

Acknowledgments

This research was supported by WCU (World Class University) program through the Korea Science and Engineering Foundation funded by the Ministry of Education, Science and Technology (R31-10008).

References

[1] K. Rohr, Towards model-based recognition of human movements in image sequences, Computer Vision, Graphics, and Image Processing: Image Understanding 59 (1) (1994) 94–115.

[2] R. Poppe, Vision-based human motion analysis: an overview, Computer Vision and Image Understanding 108 (1) (2007) 4–18.

[3] P. Felzenszwalb, D. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61 (1) (2005) 55–79.

[4] L. Sigal, M. Black, Measure locally, reason globally: Occlusion-sensitive articulated pose estimation, in: Proceedings of IEEE Computer Society

Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 2041–2048.

[5] D. Ramanan, D. A. Forsyth, A. Zisserman, Strike a pose: Tracking people by finding stylized poses, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 271–278.

[6] H. Jiang, D. R. Maritn, Global pose estimation using non-tree models, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[7] M. W. Lee, R. Nevatia, Human pose tracking in monocular sequence using multilevel structured models, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (1) (2009) 27–38.

[8] A. Agarwal, B. Triggs, Recovering 3d human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 44–58.

[9] C.-S. Lee, A. Elgammal, Modeling view and posture manifolds for tracking, in: Proceedings of IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[10] L. Raskin, M. Rudzsky, R. Rivlin, Dimensionality reduction using a gaussian process annealed particle filter for tracking and classification of articulated body motions, Computer Vision and Image Understanding 115 (4) (2011) 503–519.

[11] H.-D. Yang, S.-W. Lee, Reconstruction of 3d human body pose from stereo

image sequences based on top-down learning, Pattern Recognition 40 (11) (2007) 3120–3131.

[12] R. Li, T.-P. Tian, S. Sclaroff, M.-H. Yang, 3d human motion tracking with a coordinated mixture of factor analyzers, International Journal of Computer Vision 87 (1) (2010) 170–190.

[13] M. Ahmad, S.-W. Lee, Human action recognition using shape and clg-motion flow from multi-view image sequences, Pattern Recognition 41 (41) (2008) 2237–2252.

[14] M. Ahmad, S.-W. Lee, Variable silhouette energy image representations for recognizing human actions, Image and Vision Computing 28 (5) (2010) 81–824.

[15] M. S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking, IEEE Transactions on Signal Processing 50 (2) (2002) 174–188.

[16] C. Sminchisescu, A. Jepson, Variational mixture smoothing for non-linear dynamical systems, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 608–615.

[17] O. Bernier, P. Cheung-Mon-Chan, A. Bouguet, Fast nonparametric belief propagation for real-time stereo articulated body tracking, Computer Vision and Image Understanding 113 (1) (2009) 29–47.

[18] A. Gupta, A. Mittal, L. S. Davis, Constraint integration for efficient multiview pose estimation with self-occlusions, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (7) (2008) 493–506.

[19] P. Peursum, S. Venkatesh, W. G, A study on smoothing for particle-filtered 3d human body tracking, International Journal of Computer Vision 87 (1) (2010) 53–74.

[20] J. S. Yedidia, W. T. Freeman, Y. Weiss, Understanding belief propagation and its generalization, Tech. rep., Mitsubishi Electric Research Laboratories (January 2002).

[21] M. Park, Y. Liu, R. T. Collins, Efficient mean shift belief propagation for vision tracking, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[22] C. Bibby, I. Reid, Real-time tracking of multiple occluding objects using level sets, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 1307–1314.

[23] N. Papadakis, A. Bugeau, Tracking with occlusions via graph cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (1) (2011) 144–157.

[24] E. Sudderth, M. Mandel, W. Freeman, A. Willsky, Distributed occlusion reasoning for tracking with nonparametric belief propagation, in: Advances in Neural Information Processing Systems, 2004, pp. 1369–1376.

[25] Y. Wang, G. Mori, Multiple tree models for occlusion and spatial constraints in human pose estimation, in: Proceedings of the 10th European Conference on Computer Vision, 2008, pp. 710–724.

[26] Z. Khan, T. Balch, F. Dellaert, Mcmc-based particle filtering for tracking a

variable number of interacting targets, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1805–1819.

[27] K. Luttgens, N. Hamilton, Kinesiology: Scientific Basis of Human Motion, Madison, WI: Brown & Benchmark, 1997.

[28] L. Sigal, A. Balan, M. Black, Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, International Journal of Computer Vision 87 (1) (2010) 4–27.

[29] H.-D. Yang, S. Sclaroff, S.-W. Lee, Sign language spotting with a threshold model based on conditional random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (7) (2009) 1264–1277.

[30] X. Lan, D. Huttenlocher, Common facto models for 2d human pose recovery, in: Proceedings of IEEE International Conference on Computer Vision, 2005, pp. 470–477.

[31] M. W. Lee, I. Cohen, Proposal maps driven mcmc for estimating human body pose in static images, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 334–341.

[32] G. Mori, X. Ren, A. Efros, J. Malik, Recovering human body configurations: Combining segmentation and recognition, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 326–333.

[33] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of IEEE International Conference on Computer Vision, 2001, pp. 511–518.

[34] X. Zhang, G. Fan, Dual gait generative models for human motion estimation from a single camera, IEEE Transactions on Systems, Man, and Cybernetics 40 (4) (2010) 1034–1049.

[35] M. Schmidt, http://www.cs.ubc.ca/∼schmidtm/software/ugm.html (2007).

About the Author—NAM-GYU CHO received the B.S. in information and communication engineering from University of Incheon, Incheon, Korea and M.S. degrees in computer science and engineering from Korea University, Seoul, Korea, in 2009 and 2011, respectively. He is currently a Ph.D. candidate in the department of brain and cognitive engineering in Korea University. His research interests include computer vision, machine learning, and computational models of vision.

About the Author—Alan Yuille received the BA degree in mathematics from the University of Cambridge in 1976 and the PhD degree on theoretical physics, supervised by Professor S.W. Hawking, in 1981. He was an NATO postdoctoral research fellow studying physics at the University of Texas Austin and the Institute for Theoretical Physics at the University of California Santa Barbara in 1981/1982. He was a research scientist at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology and the Division of Applied Sciences at Harvard University from 1982 to 1988. He served as an assistant and associate professor at Harvard University until 1996. He was a senior research scientist at the Smith-Kettlewell Eye Research Institute from 1996 to 2002. In 2002, he joined the University of California Los Angeles as a full professor with a joint appointment in statistics and psychology. He received a joint appointment in computer science in 2007. His research interests include computational models of vision, mathematical models of cognition, and artificial intelligence and neural networks. He is a fellow of the IEEE.

About the Author—SEONG-WHAN LEE is the Hyundai Motor Chair Professor at Korea University, where he is the head of the Department of Brain and Cognitive Engineering and the director of the Institute for Brain and Cognitive Engineering. He received the B.S. degree in computer science and statistics from Seoul National University, Seoul, Korea, in 1984, and the M.S. and Ph.D. degrees in computer science from KAIST in 1986 and 1989, respectively. From 1989 to 1995, he was an assistant professor in the Department of Computer Science, Chungbuk National University, Cheongju, Korea. In 1995, he joined the faculty of the Department of Computer Science and Engineering, Korea University, Seoul, Korea, as a full professor. Dr. Lee was the winner of the Annual Best Student Paper Award of the Korea Information Science Society in 1986. He obtained the First Outstanding Young Researcher Award at the Second International Conference on Document Analysis and Recognition in 1993, and the First Distinguished Research Award from Chungbuk National University in 1994. He also obtained the Outstanding Research Award from the Korea Information Science Society in 1996. A Fellow of the IEEE, IAPR, and Korean Academy of Science and Technology, and IEEE SMC Society Distinguished Lecturer, he has served several professional societies as chairman or governing board member. He was the founding Co-Editor-in-Chief of the International Journal of Document Analysis and Recognition and has been an Associate Editor of several international journals; Pattern Recognition, ACM Transactions on Applied Perception, IEEE Transactions on Affective Computing, Image and Vision Computing, International Journal of Pattern Recognition and Artificial Intelligence, and International Journal of Image and Graphics, etc. He was a general or program chair of many international conferences and workshops and has also served on the program committees of numerous conferences and workshops. His research interests include pattern recognition, computer vision, and brain informatics. He has more than 250 publications in international journals and conference proceedings, and authored 10 books.