Nonflat Observation Model and Adaptive Depth Order Estimation for 3D Human Pose Tracking

Nam-Gyu Cho[†], Alan Yuille[†‡] and Seong-Whan Lee[†]
[†]Department of Brain and Cognitive Engineering, Korea University, Korea
[‡]Department of Statistics, University of California, Los Angeles, U.S.
ngcho@image.korea.ac.kr, yuille@stat.ucla.edu, swlee@image.korea.ac.kr

***Abstract*—Tracking human poses in video can be considered as to infer the information of body joints. Among various obstacles to the task, the situation that a body-part occludes another, called 'self-occlusion,' is considered one of the most challenging problems. In order to tackle this problem, it is required for a model to represent the state of self-occlusion and to efficiently compute inference, complex with a depth order among body-parts. In this paper, we propose an adaptive self-occlusion reasoning method. A Markov random field is used to represent occlusion relationship among human body parts with occlusion state variable, which represents the depth order. In order to resolve the computational complexity, inference is divided into two steps: a body pose inference step and a depth order inference step. From our experiments with the HumanEva dataset we demonstrate that the proposed method can successfully track various human body poses in an image sequence.**

***Index Terms*—Human pose tracking, Markov random field, Self-occlusion.**

I. Introduction

3D human pose tracking has great potentials for many applications such as marker-free human motion capture (which has environmental advantages compared to the marker-based system), Human Computer Interactions (HCI), Human Robot Interactions (HRI), video surveillance, and so on. This also can be explained by the fact that more than hundred papers are published on human pose estimation during the last two decades [1]. This research can be divided into two categories: discriminative approaches and generative approaches.

Discriminative approaches learn mapping between the set of features and target poses, then these mappings are stored in a database. After the learning, a set of features extracted from the input image are used to find the best matching pose from a database. Discriminative approaches can find the best matching pose quickly, and have high accuracy results when the input image is similar to the training data. However, they have poor performance when the input image is quite different from the training data [2], [3].

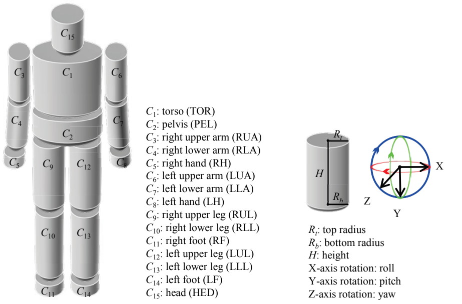
Generative approaches use graphical models, e.g., Bayesian networks or Markov networks. In these approaches, a graph node represents the state of a human body part and graph edge model the relations between parts, and these components construct probability distributions of the model and input image. These approaches are able to estimate new motions while the discriminative approaches can capture trained poses only. However, they suffer from exponentially growing computational complexity and the self-occlusion problem (a body part occludes another). Previous approaches which do not consider self-occlusion result in different parts indicating the same image area [4], [5]. Sigal et al. [6] proposed a self-occlusion reasoning method which uses occlusion-sensitive likelihood model to get a correct image representation under self-occlusion. However, although they made an important contribution to the likelihood model for self-occlusion, this approach is dependent on the type of input motion. Because they don't have a depth order estimation step, they require a large amount of additional computation, and the depth order of body parts has to be set manually for all input images. As the result of this manually set depth order, the edge of occlusion relationship, which is used for occlusion-sensitive likelihood, is constructed with respect to the input image, e.g., to track walking motion the occlusion relationship is constructed only among arms and torso, and between legs. So, these constructed occlusion relationship is only useful for walking motion.

In this paper, we propose an adaptive self-occlusion reasoning method that estimates not only body configuration but also the occlusion states (depth order) of body parts which prevents us from being dependent on the type of input motion. This is based on the following experimental observation. When the overlapping region (i.e. self-occlusion) between body parts in the image is small, then pictorial structures [5], which does not consider self-occlusions, and the self-occlusion reasoning approach [6] give similar tracking performance. But as the overlapping region gets bigger, the tracking performance of the pictorial structures approach decreases while the self-occlusion reasoning approach keeps relatively high performance, assuming manually set depth order. From this observation, we postulate that if we are able to find overlapping body parts with small overlapping regions (where pictorial structures can have fairly good estimates) then we can update the occlusion state among detected body parts and this information can be used for the next inference. Therefore, we expect that this novel scheme will lead to efficient occlusion state estimation while avoiding the combinatorial problem.

II. Adaptive Self-Occlusion Reasoning Method

Fig. 1 shows the 3D human model used in this paper. The 3D human model consists of 15 3D cylinders. Each cylinder has one of two types of DOF (Degree 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 (orientation about the X, Y, and Z axes) and C_4, C_7, C_{10}, and

978-1-4577-0121-4/11/$26.00 ©2011 IEEE



(a) 3D human model.

(b) 3D cylinder parameterization.

Fig. 1: The 3D human model and the parameterization of the 3D cylinders.

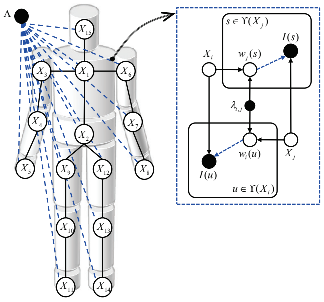


Fig. 2: The MRF for adaptive self-occlusion reasoning. Left panel: X_i $(i = 1, ..., 15)$ corresponds to each C_i in Fig. 1. Edges represent relationships: the kinematic (solid line) and occlusion (dashed blue line) relationships of body parts. Right panel: the graphical representation of the occlusion relationships encoded by dashed blue edge with respect to the occlusion state variable $\lambda_{i,j}$. See the text for the detail (Sec. II-A).

C_{13} have 1 DOF (orientation about the X axis). Cylinder C_1 has 3 additional DOF (the x, y, and z positions). The global position and orientation of the 3D human model is determined by the 6 DOF of C_1.

We formulate Markov Random Field (MRF) to represent 3D human body configuration and occlusion relationships of human body parts with occlusion state variables Λ representing depth order (Fig. 2). The probability distribution over this graph is specified by the set of potentials defined over the set of edges. These edges are specified by occlusion, kinematic, and temporal relationship among nodes. This probability distribution specifies the body configuration under self-occlusion.

A. The Structure of MRF for the Proposed Method

Notations and descriptions of the MRF for the proposed method are listed in Table I. 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: the occlusion relationships, the kinematic relationships, and the temporal relationships.

The state of X_i consists of 3D position and 3D orientation.

TABLE I: The notations used in the proposed MRF.

Notation	Description
$X = \{X_1, ..., X_{15}\}$	the set of nodes for body parts
$\mathbf{x}_i = (x, y, z)$	position of X_i in 3D space
$\boldsymbol{\theta}_i = (\theta_x, \theta_z, \theta_z)$	orientation of X_i in 3D space
$\Upsilon(X_i)$	the set of pixels in the area in the image where X_i projects to
$W_i = \{w_i(u)\}, (u \in \Upsilon(X_i))$	the set of visibility variables of pixel u's.
$\Lambda = \{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_{15}\}$	the set of occlusion state variables
$\boldsymbol{\lambda}_i = (\lambda_{i,1}, ...\lambda_{i,15}), \lambda_{i,i} = 0$	the set of occlusion state variables between node X_i and the others
$E = (E_K, E_{O\|\Lambda}, E_T)$	the 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

The scale of cylinder is determined by the z value of 3D position (scaled orthographic projection [7]); the larger the z value of $\mathbf{x}_i$, the closer it is to the camera.

$w_i(u)$ represents the visibility of pixel u generated by the X_i and has binary state like following,

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

where the value of this state variable is determined by the state of the occluders of X_i with respect to the state of $\boldsymbol{\lambda_i}$. The occlusion state variable $\boldsymbol{\lambda_i}$ is only defined on 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 none of two body parts occludes the other, state 2 represents body part X_i occludes body part X_j, and state 3 is when body part X_i is occluded by body part X_j.

B. The Probability Distribution

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

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

$$\phi_i(I, X_i; \boldsymbol{\lambda}_i) = \phi_i^C(I, X_i; \boldsymbol{\lambda}_i) + \phi_i^E(I, X_i; \boldsymbol{\lambda}_i), \tag{2}$$

where ϕ_i^C is the observation potential for the color cue and ϕ_i^E is for the edge cue. A 3D cylinder is generated from the state of X_i (i.e. a set of pixels are generated). We modified the occlusion-sensitive likelihood model [6] for the observation potential with respect to the occlusion state variable Λ. A critical issue of the occlusion-sensitive likelihood model is how to find the configuration of W_i, the set of visibility variables about X_i. The depth order for current image was assumed to be known in [6]. However, using the proposed occlusion state variable Λ, the configuration of W_i can be calculated deterministically. For example, in Fig. 3, X_{RLA} is occluded by X_{TOR}. In this case, let's assume that ϕ_{TOR} is calculated in advance. The configuration of W_{RLA} is determined by the calculating overlapping region of X_{RLA} and X_{TOR}. Therefore, W_i can be represented as,

$$W_{RLA} = \{w_{RLA}(u'), w_{RLA}(u)\}, \tag{3}$$

where $w_{RLA}(u')$=0 for $u' \in \Upsilon'(X_{RLA})$ where $\Upsilon'(X_{RLA}) = (\Upsilon(X_{TOR}) \cap \Upsilon(X_{RLA}))$. And $w_{RLA}(u) = 1$ for $u \in (\Upsilon(X_{RLA}) - \Upsilon'(X_{RLA}))$. This leads to calculating the observation potential separately. The observation potential for the color cue is formulated as follows:

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

where the first term is for the visible area, and the second terms is for the occluded area. The visible term is formulated as,

$$\phi_i^{C_V}(I, X_i; \boldsymbol{\lambda}_i) = \prod_{u \in (\Upsilon(X_i) - \Upsilon'(X_i))} p_C(I_u), \tag{5}$$

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

$$p_C(I_u) = \frac{p(I_u|\text{foreground})}{p(I_u|\text{background})}, \tag{6}$$

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

$$\phi_i^{C_{OC}}(I, X_i; \boldsymbol{\lambda}_i) = \prod_{u' \in \Upsilon'(X_i)} [z_i(I_{u'}) + (1 - z_i(I_{u'}))p_C(I_{u'})], \tag{7}$$

where $z_i(I_{u'})$ is the probability of pixel u' being explained by occluder's pixel probability. For example, again in Fig. 3, there is no problem for calculating ϕ_{TOR} for both the gray visible and the yellow overlapping regions (previously, we assumed that ϕ_{TOR} is calculated in advance). But, a problem occurs in the occluded yellow region for calculating ϕ_{RLA}. So, this potential is approximated by weighting down the pixel probability $p_C(I_{u'})$ of X_{RLA} by the pixel probability of X_{TOR}. Therefore, $z_{RLA}(I_{u'})$ is the normalized pixel probability of X_{TOR} about pixel u' (normalized with respect to the whole states of X_{TOR}). $\phi_i^E(I, X_i; \boldsymbol{\lambda}_i)$ is calculated in the same way as $\phi_i^C(I, X_i; \boldsymbol{\lambda}_i)$ where $p_E(I_u)$ is calculated from chamfer distance transform.

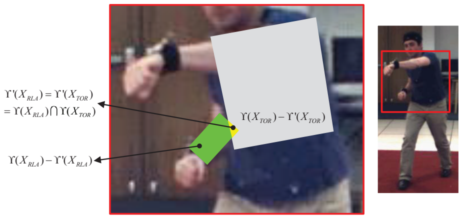

$\Upsilon'(X_{RLA}) = \Upsilon'(X_{TOR})$
$= \Upsilon(X_{RLA}) \cap \Upsilon(X_{TOR})$
$\Upsilon(X_{TOR}) - \Upsilon'(X_{TOR})$
$\Upsilon(X_{RLA}) - \Upsilon'(X_{RLA})$

Fig. 3: An example of self-occlusion between TOR and RLA where RLA is occluded by TOR. Green and gray region are non-overlapping region of X_{TOR} and X_{RLA} respectively. The yellow region is the overlapping region of X_{TOR} and X_{RLA}.

2) Kinematic Potential: We model the kinematic relationship between two adjacent parts using Kinesiology, which defines the Range of Motion (ROM) of human joints [8]. We use ROM to approximate the possible range of orientation of adjacent body parts in 3D space. This kinematic relationship of position and orientation of two adjacent body parts is formulated as,

$$\psi_{ij}^K(X_i, X_j) = N(d(\mathbf{x}_i, \mathbf{x}_j); 0, \sigma_K) f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j), \tag{8}$$

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 (for this, $i < j$). $N()$ is the normal distribution with mean zero and standard deviation σ_K to allow adjacent body parts to be linked loosely and $f()$ has value 0 when the rotation relation (ROM) is violated, and 1 for vice versa.

3) Temporal Potential: The temporal relationship of a part between two consecutive time steps $t-1$ and t follows a Gaussian distribution as follows:

$$\psi_i^T(X_i^t, X_i^{t-1}) = p(X_i^t - X_i^{t-1}; \boldsymbol{\mu}_i, \Sigma_i), \tag{9}$$

where $\boldsymbol{\mu}_i$ is a dynamics of X_i at previous time step and Σ_i is a diagonal matrix with the diagonal elements are same as $|\boldsymbol{\mu}_i|$.

The posterior distribution of model X given all input images up to current time step τ and occlusion state variable Λ is an example of a pairwise MRF over time steps from 1 to τ [9],

$$\begin{aligned} p(X^\tau | I^{1:\tau}; \Lambda^{1:\tau}) = \frac{1}{Z} \exp\{ - \sum_{i \in X^{1:\tau}} \phi_i^C(I, X_i; \boldsymbol{\lambda}_i) \\ - \sum_{ij \in E_K^{1:\tau}} \psi_{ij}^K(X_i, X_j) - \sum_{i \in E_T^{1:\tau}, t \in 1:\tau} \psi_i^T(X_i^t, X_i^{t-1}) \}, \end{aligned} \tag{10}$$

where the observation potential and the kinematic potential are calculated over whole time step. Therefore, only the temporal potential contains on explicit time index t.

C. Inference: Body Configuration Estimation

The goal of inference is to obtain the best state of X and Λ from the conditional probability distribution in (10) at time step t. In this paper, we use two steps to estimate 3D body configuration and occlusion state variable separately. Let's 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 body configuration

$\hat{X}^t$ from input image at current time step t. This is formulated as follows:

$$\hat{X}^t = \underset{X^t}{\arg\max}\, p(X^t | I^{1:t}; \hat{\Lambda}^{t-1}). \tag{11}$$

In order to calculate this inference efficiently, we use the Belief Propagation (BP) [10] algorithm. The BP uses local messages that sum up all the possible probabilities about neighbor nodes with regard to a state of node. In this paper, we conducted the BP over two consecutive time steps $t-1$ and t using UGM (Undirected Graphical Model) toolbox [11]. In order to find the most likely body configuration X^t, we adapted 2 annealing steps [12] for inference.

D. Occlusion State Estimation

To estimate the occlusion state among 15 body parts, we have to consider $3^{15} (\simeq 10^7)$ possible combinations of occlusion state variable Λ. And, since the most time consuming process of the inference is the observation potential, this combinatorial problem makes the complexity even bigger. In this paper, we propose a novel occlusion state estimation method based on our experimental observation (see Sec. I). In order to find overlapping (occluding) body parts, we first define a criterion for detecting overlapping body parts as,

$$\nu_{i,j} = \begin{cases} 1, & \text{if } max\left(\dfrac{OR_{i,j}}{\Upsilon(\hat{X}_i^t)}, \dfrac{OR_{i,j}}{\Upsilon(\hat{X}_j^t)}\right) \geq T_0 \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where $OR_{i,j}$ is the overlapping region between $\Upsilon(\hat{X}_i^t)$ and $\Upsilon(\hat{X}_j^t)$. T_0 is a threshold set as 0.15, determined by experiments. If the value of $\nu_{i,j}$ is set to 0, the value of $\lambda_{i,j}^t$ is set to zero. Otherwise, $\lambda_{i,j}^t$ is estimated using the following equation. This criterion was also used to find the overlapping areas between two hands [13]. After the detection step, occlusion state estimation is conducted only for detected body parts X_i and X_j. So, one of occlusion state 2 and 3 is estimated as a state that has higher observation potential value as follows:

$$\hat{\lambda}_{i,j}^t = \underset{\lambda_{i,j} \in \{2,3\}}{\arg\max}\, \phi(I^t, \hat{X}_i^t; \lambda_{i,j}), \tag{13}$$

where $\hat{X}_i^t$ is the estimate of X_i^t at previous step. If $\nu_{i,j}$ has value 0 then $\hat{\lambda}_{i,j}$ is assigned as 1.

E. Proposals

In human body pose estimation, strong priors improves the robustness of performance, but also has limitations [14]. Robust body part detectors, e.g., for head, torso, and limbs, make the pose estimation task easier. This reduces the search space ,but is not reliable all the time mainly due to the image noise and self-occlusions [15]. In this paper, we construct proposals for the head and torso: face detector [16] and head-shoulder contour detector for torso [15]. 50 samples of each part that have the most likely states are selected to build proposals.

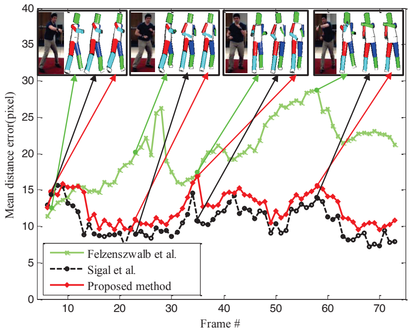


Fig. 4: Tracking performance for boxing motion. Felzenszwalb et al. [5] gradually loses arm parts while Sigal et al. [6] and the proposed method track with fairly good performance.

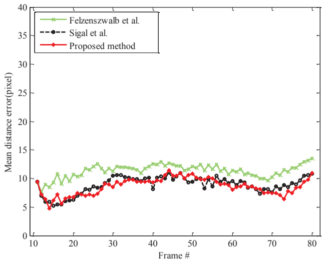


Fig. 5: Tracking performance for gesture motion. Since gesture motion contains small self-occlusions, the three methods show similar performance.

III. Experimental Results and Analysis

The HumanEva dataset [17] is used for evaluation. It contains 6 different motions with ground truth. For the evaluation, however, we excepted the values of torso because these values were set to NaN (Not a Number) in the ground truth data. We compared three methods: pictorial structures [5], self-occlusion reasoning [6] with known depth orders, and the proposed adaptive self-occlusion reasoning method with 74 frames of boxing motion and 69 frames of gesture motion. To make a fair comparison, we set all methods to have 15 nodes, the same as the proposed method. We also initialized the body configuration state, occlusion state parameter (especially for the Sigal et al. [6] and the proposed method) and the 3D human model for the three methods manually. And in order to evaluate the performance of occlusion state parameter estimation, we manually specified occlusion state for every single image of three input sequences. On the average, it took 3 minutes. These manually set occlusion states are used for Sigal et al. [6]. The method was implemented in MATLAB R2009a, and the experiments were conducted on a desktop PC (Intel core 2 Quad 2.66GHz CPU, 4GB RAM, and Windows 7 operating system).

A. Pose Tracking

The tracking performance of three methods are evaluated by an Euclidean distance error between ground truth and estimate. Although we estimated the 3D body configuration, it is hard to compare our estimates and ground truth directly because

TABLE II: Mean distance errors of three methods.

Motion	Felzenszwalb et al.	Sigal et al.	Proposed method
Boxing	20.48	10.58	12.48
Gesture	11.32	9.10	8.61
Mean	15.9	9.84	10.54

*unit: pixel

TABLE III: Mean of occlusion state estimation error.

Motion	Boxing	Gesture
Mean	7.71	3.50

*unit: %

they have different coordinate system. Thus, we converted both estimates and ground truth values into the same 2D coordinate system. E^t_{BC}, an error of the estimation of body configuration at time step t, is calculated by the following equation,

$$E^t_{BC} = \frac{\sum_{i=1}^{15} d(X^t_{est,i}, X^t_{gnd,i})}{15}, \qquad (14)$$

where $d(X^t_{est,i}, X^t_{gnd,i})$ is an Euclidean distance, $X^t_{est,i}$ and $X^t_{gnd,i}$ are the estimate and the ground truth of individual body part X_i at time step t respectively.

Fig. 4 and Fig. 5 illustrate the tracking performance of three methods for boxing and gesture motion respectively. The proposed method is able to track the human body pose from an image sequence with adaptive estimation of depth orders while Sigal et al. [6] is only able to track with when depth orders are given.

The mean distance errors are represented in Table II. Felzenszwalb et al. [5] has the largest mean distance error for both motions and Sigal et al. [6] has the smallest error for boxing motion. The proposed method has a little larger than Sigal et al. [6] and the smallest error for gesture motion. In particular, the proposed method obtained this result without known depth order. We insist that this proves that the proposed method is robust to not only the self-occlusion but also the different types of motions.

B. Occlusion state estimation

The performance of the occlusion estimation is evaluated by the mean error. The error at time step t is calculated as follows,

$$E^t_{ose} = \frac{UpperTriangular(\Lambda^t_{gnd} - \hat{\Lambda}^t)}{K}, \qquad (15)$$

where Λ^t_{gnd} is the ground truth of occlusion state at time step t and $\hat{\Lambda}^t$ is the estimate by the proposed method. $UpperTriangular(\Lambda)$ is a function that returns the upper triangular matrix of Λ and K is the total number of elements in upper triangular matrix of Λ ($K = 120$). Table III shows the mean of occlusion state estimation error for two input motions.

IV. Conclusion

In this paper, we proposed the adaptive self-occlusion reasoning method for 3D human pose tracking. While the previous self-occlusion reasoning approach [6] was able to track only if the depth orders for the whole input images are given our method successfully tracked without manually set depth orders. The proposed method infers state variables efficiently due to separating estimation procedure into body configuration estimation and occlusion state parameter estimation. This leads to efficient estimation with a small amount of additional time cost. Experimental results have shown that the proposed methods successfully tracks the 3D human pose in the presence of self-occlusion.

Acknowledgment

This research was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-10008-0).

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

[3] 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.

[4] D. Ramanan, D. A. Forsyth, and 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.

[5] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[6] L. Sigal and 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.

[7] K. Riley, *Mathematical Methods for Physics and Engineering*. Cambridge University Press, 2006.

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

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

[10] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalization," Mitsubishi Electric Research Laboratories, Tech. Rep., January 2002.

[11] M. Schmidt, "http://www.cs.ubc.ca/∼schmidtm/software/ugm.html," 2007.

[12] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 185–205, 2004.

[13] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1264–1277, 2009.

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

[15] M. W. Lee and 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.

[16] P. Viola and 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.

[17] L. Sigal, A. Balan, and M. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1, pp. 4–27, 2010.