2013 IEEE RO-MAN: The 22nd IEEE International Symposium on Robot and Human Interactive Communication
Gyeongju, Korea, August 26-29, 2013

Incorporating Global and Local Observation Models for Human Pose Tracking

Nam-Gyu Cho and Seong-Whan Lee

Abstract— **Tracking human pose is attractive to many applications such as Human Robot Interface (HRI), motion capture system, video surveillance, action recognition, etc. Though various methods were introduced during last decades, including both color and depth camera based, it is still considered that feature sets for them are not discriminative enough. In this paper, we propose a human pose tracking method based on a graphical model which incorporates global and local feature sets including Histogram of Oriented Gradients (HOG) and color distribution. HumanEva-I dataset is used for testing effectiveness of the proposed method.**

I. INTRODUCTION

Tracking human pose is attractive to many applications such as Human Robot Interface (HRI), motion capture system, video surveillance, action recognition, etc. Particularly, video surveillance and action recognition can be greatly helped by the pose tracking result or vice versa [1]–[4]. However, during last decades, various methods were introduced for human pose tracking problem from a monocular image sequence [5]. There are lots of challenges such as various appearances and poses of human, environmental noises including illumination and self-occlusion – where one body part is occluded by another one. Some methods exploited feature information – e.g., silhouette, edge, and texture – from an image, but their feature set don't have enough discriminative power, particularly under self-occlusion [6]–[9]. Defining occlusion state [10], [11] or using multiple camera information help handling self-occlusion [12], [13]. But they have limitations of features for describing human pose.

In this paper, we propose a human pose tracking method based on a graphical model which incorporates global and local feature set including Histogram of Oriented Gradients (HOG) and color distribution. Our method follows the framework of [11]. However, our method is different in terms of features to maximize discriminative power.

The rest of this paper is organized as follows: Section II reviews occlusion reasoning method [11]. Section III provides the proposed discriminative feature set for a human pose tracking. Section IV presents and discusses experimental results. Section V concludes this paper.

II. ADAPTIVE OCCLUSION STATE ESTIMATION METHOD

3D human model consists of 15 3D cylinders (Fig. 1(a)) is used. C_2, C_3, C_5, C_6, C_8, C_9, C_{11}, C_{12}, C_{14} and C_{15} have 3 DOF (Degrees of Freedom) – 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. C_1 has 6 DOF (the x, y, and z positions and rotation about each axis). Therefore, the global position and orientation of the 3D human model is determined by the 6 DOFs of C_1.

Nam-Gyu Cho and Seong-Whan Lee are with Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea {ngcho, swlee}@image.korea.ac.kr

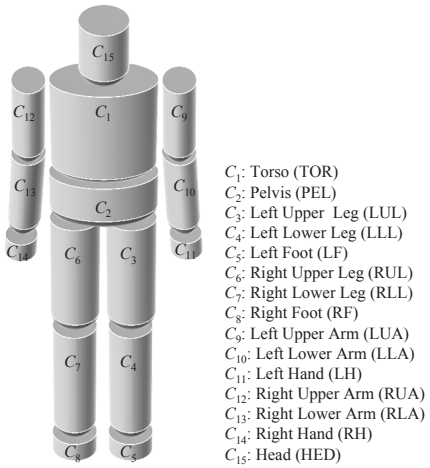
TABLE I
THE NOTATIONS USED IN THIS PAPER.

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

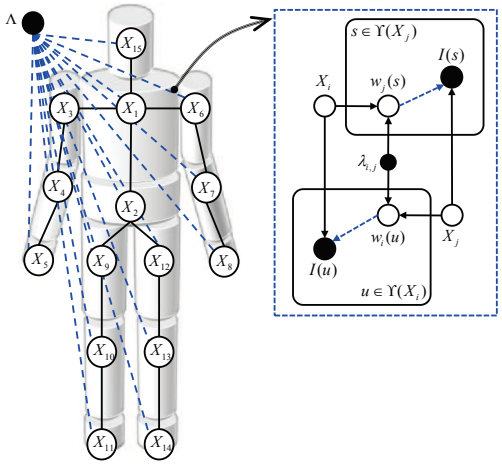
3D human pose tracking is conducted using a graphical model with state variables X for 3D body configuration and Λ for occlusion relationship (please see Table I for notations in this paper). The goal of human pose tracking is to determine the posterior distribution $p(X^\tau|I^{1:\tau})$ for the body configuration X^τ at time τ, given all input images $I^{1:\tau} = \{I^1, ..., I^\tau\}$ [14], and the distribution can be formulated with set of potentials the graph [15] as follows,

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

where Z is a normalization constant.



(a) 3D human model.



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

Fig. 1. The 3D human model and graphical model used in this paper. Semantically, C_i corresponds to X_i $(i = 1, ..., 15)$.

A. Structure of the Graphical Model

Formally, a graph is $G = (V, E)$ where V is the set of nodes and E is the set of edges. The graph has state variables X, W and Λ. The edges encode relationships (occlusion, kinematic, and temporal) and these relationships are modeled as the set of potentials. The occlusion relationship $E_{O|\Lambda}$ is formed over 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. The kinematic relationship E_K is formed over adjacent nodes. The temporal relationship E_T is formed over X_i^t and X_i^{t+1}.

B. 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 the X_i and it has a binary state defined by,

$$w_i(u) = \begin{cases} 0, & \text{if pixel } u \text{ is occluded} \\ 1, & \text{if pixel } u \text{ is not occluded,} \end{cases} \quad (2)$$

where the value of this state variable is determined by the state of the occluders of X_i with respect to the state of $\boldsymbol{\lambda_i}$. As illustrated in Fig. 2, $\lambda_{i,j}$ is only defined between different body parts (i.e. $\lambda_{i,i} = 0$). 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.

C. The Potentials

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. Two image cues are used – color and edges – to calculate the observation potential as follows:

$$\phi_i(I, X_i; \boldsymbol{\lambda}_i) = \phi_i^C(I, X_i; \boldsymbol{\lambda}_i) + \phi_i^E(I, X_i; \boldsymbol{\lambda}_i) + \phi_i^{global}(I, X_i), \quad (3)$$

where ϕ_i^C is the observation potential for the color cue and ϕ_i^E is the observation potential for the edge cue. ϕ_i^{global} is used to describe global property of an observation (please see Section III for a detailed description) which is proposed in this paper. The configuration of W_i can be calculated deterministically. W_i can be represented as,

$$W_i = \{w_i(u'), w_i(u)\}, \quad (4)$$

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 calculation of the observation potential. The observation potential for the color cue is formulated as follows:

$$\phi_i^C(I, X_i; \boldsymbol{\lambda}_i) = \phi_i^{C_{visible}}(I, X_i; \boldsymbol{\lambda}_i) + \phi_i^{C_{occluded}}(I, X_i; \boldsymbol{\lambda}_i), \quad (5)$$

where the first and second term of the right hand side are for the visible and occluded area respectively. The visible term is formulated as,

$$\phi_i^{C_{visible}}(I, X_i; \boldsymbol{\lambda}_i) = \prod_{u \in (\Upsilon(X_i) - \Upsilon'(X_i))} p_C(I_u), \quad (6)$$

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

$$p_C(I_u) = \frac{p(I_u|\text{foreground})}{p(I_u|\text{background})}, \quad (7)$$

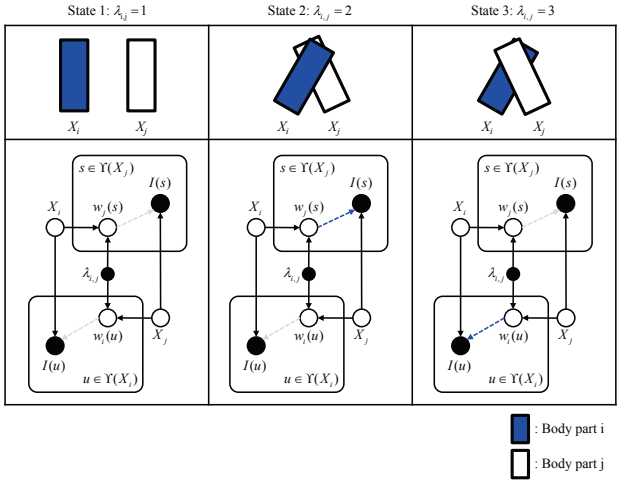


Fig. 2. Definition of three occlusion states between two body parts. The topology of $E_{O|\Lambda}$ (dashed edges in blue) changes as the value of $\lambda_{i,j}$ changes.

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,

$$\phi_i^{C_{occluded}}(I, X_i; \boldsymbol{\lambda}_i) = \prod_{u' \in \Upsilon'(X_i)} [z_i(I_{u'}) + (1 - z_i(I_{u'}))p_C(I_{u'})], \quad (8)$$

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

$$z_i(I_{u'}) = \frac{1}{N_O} \sum_{\forall X_j \text{ s.t. } \lambda_{i,j}=4} \phi_j^C(I(u'), X_j^s; \boldsymbol{\lambda}_i), \quad (9)$$

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

kinematic relationship of the position and orientation of two adjacent body parts is formulated as Gaussian distribution with $\mu_K(=0)$ and σ_K to allow adjacent body parts to be loosely linked. The distribution of this relationship is learnt from the HumanEva dataset [16].

Temporal potential models the temporal relationship of a part between two consecutive time steps $t-1$ and t as a Gaussian distribution with a mean as $\boldsymbol{\mu}_i$, the dynamics of X_i at the previous time step, and a standard deviation as Σ_i which is a diagonal matrix with diagonal elements identical to $|\boldsymbol{\mu}_i|$.

D. Inference

Inference is conducted by dividing it into two steps to estimate the 3D body configuration and occlusion state variable separately. This uses the assumption 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,

$$\hat{X}^t = \arg\max_{X^t} p(X^t | I^{1:t}; \hat{\Lambda}^{t-1}). \quad (10)$$

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. We use SIR (Sequential Importance Resampling) principle to represent posterior distribution (approximately equals to the belief) of each body part by a set of N random samples with corresponding weights. In order to find the modes of the posterior distribution better, we iterate SIR steps 2 times for body configuration inference at each time step.

There are 3^{14} ($\simeq 10^5$) possible combinations of occlusion state variable Λ^t and this increases the complexity further. In order to solve this problem, we first find overlapping (occluding) body parts with a criterion – similar to [17] – for detecting overlapping body parts as follows,

$$\nu_{i,j} = \begin{cases} 1, \text{if } max \\ \left(\dfrac{\Upsilon(\hat{X}_i^t) \cap \Upsilon(\hat{X}_j^t)}{\Upsilon(\hat{X}_i^t)}, \dfrac{\Upsilon(\hat{X}_i^t) \cap \Upsilon(\hat{X}_j^t)}{\Upsilon(\hat{X}_j^t)} \right) \geq T_0 \\ 0, \text{otherwise} \end{cases} \quad (11)$$

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,

$$\hat{\lambda}_{i,j}^t = \arg\max_{\lambda_{i,j} \in \{2,3\}} \phi(I^t, \hat{X}_i^t; \lambda_{i,j}), \quad (12)$$

where $\hat{X}_i^t$ is the estimate of X_i^t from the previous step.

Strong priors improve the performance robustness, but also have limitations [18]. Robust body part detectors reduce the search space, but it is not always reliable, which is mainly due to the image noise and self-occlusions [19]. In this paper, proposals are built for the head and torso: a face detector [20] and a head-shoulder contour detector for the torso [19]. 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.

III. PROPOSED GLOBAL OBSERVATION MODEL

In this paper, we exploited a set of features that are robust to noises due to occlusion and ambiguous background. As in equation (3), global observation potential is modeled as follows,

$$\phi_i^{global}(I, X_i) = \phi_i^{shape}(I, X_i) + \phi_i^{color}(I, X_i). \quad (13)$$

where $\phi_i^{shape}(I, X_i)$ and $\phi_i^{color}(I, X_i)$ measures shape and color histogram of state X_i respectively. Each term is described by following two subsections.

A. Histogram of Oriented Gradients (HOG)

HOG was proposed to detect human object from clutter scene [21]. This has been widely used for object detection [22] and object part detection [3], [23], [24]. HOG calculates a histogram of gradients within a given cell and normalizes

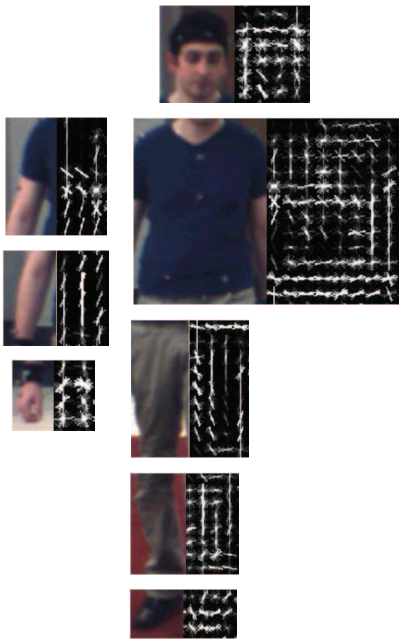


Fig. 3. Examples of learned part templates of human body. Each pair represents a body part (left side) and corresponding template (right side).

them. This gives us an invariant feature of a cell so that it can be used as a kind of global descriptor of current part observation in order to reduce an ambiguity caused by occlusion or other noises while other observation potentials – the first and second terms of the right hand side of equation (5) – focus on specific states, i.e. occluded or not.

In this paper, we learn part templates according to their 3D configuration from the training dataset [16]. Firstly, we analyze the range of rotation of each joint and quantize them according to the range of each joint. Then we learn part template using linear SVM same as [23]. Fig. 3 shows examples of learned part templates.

Potential of the shape is calculated as,

$$\phi_i^{shape}(I, X_i) = \chi^2(X_i, X_{\text{shape template}}). \quad (14)$$

where $\chi^2(,)$ is Chi-square distance, and $X_{\text{part template}}$ is template model of the target part.

B. Color Distribution

We learned color distribution of parts in the CIELab color space. This has much larger space than RGB or CMYK color model so that we can reduce ambiguities caused by environmental situation of the dataset. Distribution of each part is learned separately to maximize discriminative ability. We learned distributions using images of subject 2 of the dataset. It results six distribution models for head, torso, arms, hands, legs, and feet.

Potential of the color distribution is calculated as,

$$\phi_i^{color}(I, X_i) = \chi^2(X_i, X_{\text{color template}}). \quad (15)$$

where $X_{\text{color template}}$ is the color histogram of the target part. This is learned by counting frequencies components of the color space. We quantized each of L, a, and b component into 8 bins so that we have the feature vector in 24 dimensional space. Finally, we normalize $X_{\text{color template}}$ by its size.

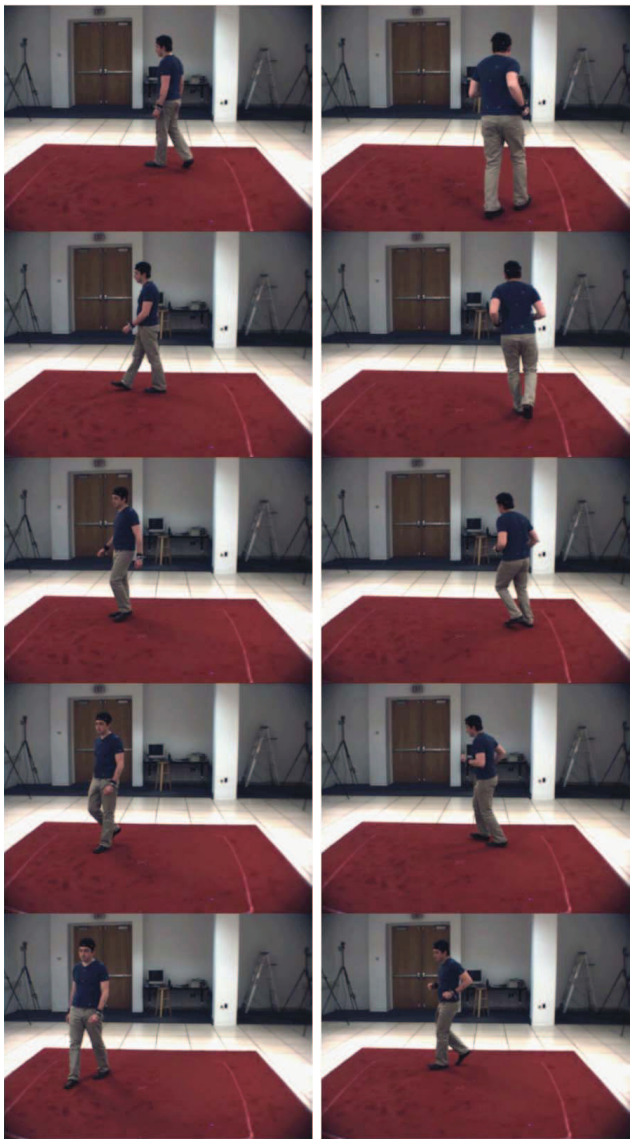


Fig. 4. Examples of input images of Walking (left panel) and Jogging (right panel).

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we demonstrate efficiency of the proposed method in terms of tracking accuracy by measuring Euclidean distance between the ground truth and estimated joints. We use Walking and Jogging motions of the HumanEva-I dataset [16] for evaluation (see Fig 4).

We compare five methods: (1) Pictorial Structure (PS) [6], (2) Self-occlusion Reasoning (SR) [10], (3) Mixture of Factor Analyzers (FMA) [25], (4) Adaptive Occlusion Estimation (AOE) [11], and (5) the proposed method. FMA tracks 3D human pose on a manifold space with multi-view information (camera C1-C3) while AOE and the proposed method use only single view (camera C1) information. During the

TABLE II

THE TRACKING ERRORS IN MILLIMETERS OVER 150 FRAMES.

Method	Walking	Jogging	Mean
PS [6]	230.66 (209.85)	113.99 (100.85)	153.39 (100.85)
SR [10]	144.74 (144.11)	164.02 (144.11)	120.97 (69.02)
FMA [25]	68.67 (24.66)	72.14 (29.62)	69.30 (29.62)
AOE [11]	88.03 (42.71)	53.58 (17.95)	70.81 (30.33)
Our method	90.42 (29.48)	51.30 (21.11)	70.86 (25.26)

experiment, initialization for the first input image were done manually for AOE and the proposed method.

In Table II, the mean and the standard deviation of tracking error of the four 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 [25]. Same as [11], we don't count the error from completely occluded parts (for PS, SR, AOE, and the proposed method).

Overall, the proposed method outperforms all the other methods for Jogging motion. Particularly, the accuracy increased when it compared to AOE. We argue that it is due to our observation model which incorporates invariant shape and color histogram features. However, for Walking motion, the performance decreased by 2 then AOE. We think that this is due to the nature of the target motion. Since Walking motion is less complex than Jogging – i.e. when human walks, he swings his arms slowly when it compared to Jogging, so Jogging gives more ambiguous image observations then Walking. The complexity of motion can be supported by the analysis done by [11]. They analyzed the motion complexity in terms of Frame interval per Occlusion state Change. Therefore, the proposed observation model could resolve ambiguities from Jogging and is found to be less effective for Walking. FMA give the best performance because it tracks human pose from multi-view information so that it can exploits more useful feature.

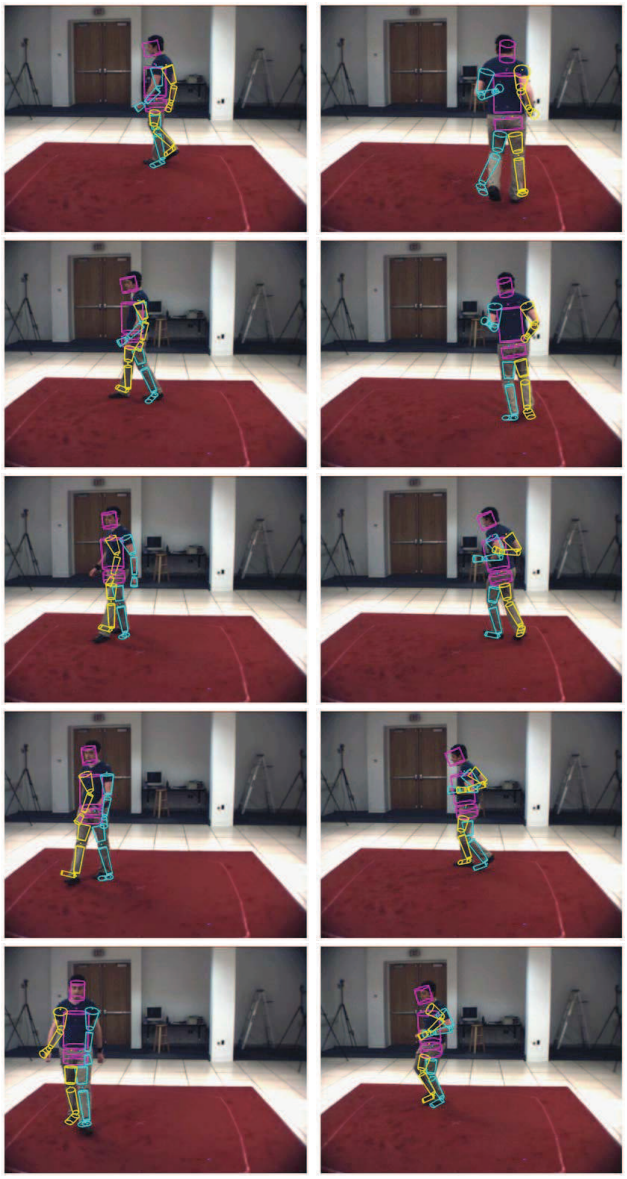


Fig. 5. Tracking results of the proposed method for Walking (left panel) and Jogging (right panel) motions.

V. CONCLUSION

In this paper, we proposed a human pose tracking method based on a graphical model which incorporates global and local feature set including Histogram of Oriented Gradients (HOG) and color distribution. Our method follows the framework of [11]. However, our method is different in terms of features to maximize discriminative power. HOG enabled the model to have richer representation for part observation. By incorporating the global shape and color model, we could deal with ambiguities caused by self-occlusions and noises from environmental situation.

We tested the proposed method on two motions – Walking and Jogging which have a high motion complexity – of challenging benchmark dataset (HumanEva-I). The results proved that the proposed method is more robust than previous methods. The tracking accuracy increased for Jogging while slightly decreased for Walking. As a future work, we conjecture that analysis of low level properties of image observations is necessary so as to use the global observation model for smoothing ambiguities while the local observation model deals with detailed image features under self-occlusions or noises.

ACKNOWLEDGMENT

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) and also supported by the Implementation of Technologies for Identification, Behavior, and Location of Human based on Sensor Network Fusion Program through the Ministry of Knowledge Economy (Grant Number: 10041629).

REFERENCES

[1] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2030–2037.

[2] G. F. Angela Yao, Juergen Gall and L. V. Gool, "Does human action recognition benefit from pose estimation?" in *Proceedings of the British Machine Vision Conference*, 2011, pp. 1–11.

[3] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1385–1392.

[4] A. Yao, J. Gall, and L. Gool, "Coupled action recognition and pose estimation from multiple views," *International Journal of Computer Vision*, no. 1, pp. 16–37, 2012.

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

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

[7] 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.

[8] H. Jiang and 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.

[9] M. W. Lee and R. Nevatia, "Human pose tracking in monocular sequence using multilevel structured models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 27–38, 2009.

[10] 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.

[11] N.-G. Cho, A. Yuille, and S.-W. Lee, "Adaptive occlusion state estimation for human pose tracking under self-occlusions," *Pattern Recognition*, vol. 46, no. 3, pp. 649–661, 2013.

[12] O. Bernier, P. Cheung-Mon-Chan, and A. Bouguet, "Fast nonparametric belief propagation for real-time stereo articulated body tracking," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 29–47, 2009.

[13] H.-D. Yang and S.-W. Lee, "Reconstruction of 3d human body pose from stereo image sequences based on top-down learning," *Pattern Recognition*, vol. 40, no. 11, pp. 3120–3131, 2007.

[14] Z. Khan, T. Balch, and F. Dellaert, "Mcmc-based particle filtering for tracking a variable number of interacting targets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1805–1819, 2005.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[22] Y. Chen, L. Wan, L. Zhu, R. Fergus, and A. Yuille, in *The PASCAL Visual Object Classes Challenge Workshop*, 2011.

[23] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 1365–1372.

[24] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 9, pp. 886–893, 2010.

[25] R. Li, T.-P. Tian, S. Sclaroff, and M.-H. Yang, "3d human motion tracking with a coordinated mixture of factor analyzers," *International Journal of Computer and Vision*, vol. 87, no. 170-190, pp. 1034–1049, 2010.