2012 IEEE International Conference on Systems, Man, and Cybernetics
October 14-17, 2012, COEX, Seoul, Korea

# Self-Occlusion Robust 3D Human Pose Tracking from Monocular Image Sequence

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*—Pose tracking technique has great potential for many applications such as marker-free human motion capture system, Human Computer Interactions (HCI), and video surveillance. Though many methods are introduced during last decades, self-occlusion – one body part is occluded by another one – is still considered one of the most difficult problems for 3D human pose tracking. In this paper, we propose a self-occlusion state estimation method. A MRF (Markov Random Field) is used to model the occlusion state which represents the pairwise depth order between two human body parts. A novel estimation method is proposed to infer a body pose and an occlusion state separately. HumanEva dataset is used for testing the proposed method. In order to evaluate and quantify how often the occlusion state changes, we label the ground truth of occlusion state.**

***Index Terms*—3D human pose tracking, Self-occlusion, Motion analysis**

Though many methods are introduced during last decades, self-occlusion – where one body part is occluded by another one – is still considered one of the most difficult problems for 3D human pose tracking [1]. Self-occlusion occurs when a human rotates with respect to the camera view point or performs a dynamic motion such as boxing and jogging. This makes it difficult for many methods to get correct feature information – e.g., silhouette, edge, and texture – from an image [2], [3], [4], [5]. Using multiple camera information helps handling self-occlusion by summing up all possible informations from each camera [6], [7]. In other field, e.g., action recognition, multiple camera information helps performance [8]. But, it has an environmental limitation [9]. Self-occlusion Reasoning (SR) method is proposed to cope with the ambiguity problem due to self-occlusion. If the depth order is known and fixed, SR can measure the likelihood of self-occluded body part approximately [10]. However, there several limitations: the method requires a known and unchanging depth order, e.g., a walking motion recorded at the lateral view, and this can't always be common situation for most human motions.

In this paper, we propose a self-occlusion robust human pose tracking method. We introduce a variable of the occlusion state modeled by three states. In order to reduce the computational problem for estimating occlusion states, we propose a novel inference scheme that estimates body configuration and occlusion states separately.

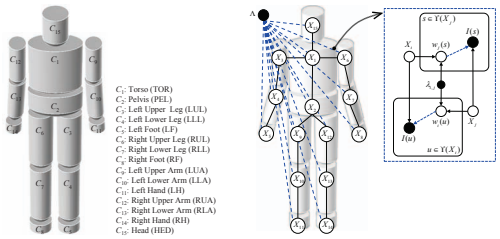
## I. The Adaptive Occlusion State Estimation Method

3D human model used in this paper consists of 15 3D cylinders (Fig. 1(a)). $C_1$, $C_2$, $C_3$, $C_5$, $C_6$, $C_8$, $C_9$, $C_{11}$, $C_{12}$, $C_{14}$ and $C_{15}$ have 3 DOF (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 3 additional DOFs (the x, y, and z positions). Therefore, the global position and orientation of the 3D human model is determined by the 6 DOFs of $C_1$.

We track the 3D human pose using a Markov Random Field (MRF) with state variables $X$ and $\Lambda$ which means the body

TABLE I
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 |
| $\phi_i$ | potential of observation |
| $\psi_{ij}^K$ | potential of kinematic relationship |
| $\psi_i^T$ | potential of temporal relationship |




(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.

configuration and occlusion relationship respectively (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 $X^\tau$ at time $\tau$, given all input images $I^{1:\tau} = \{I^1, ..., I^\tau\}$ [11], and the distribution can be formulated with set of potentials over MRF [12] as follows,

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

where $Z$ is a normalization constant.

*1)* **The Structure of the MRF***:* Formally, the MRF is a graph $G = (V, E)$ where $V$ is the set of nodes and $E$ is the set of edges. The graph has state variables $X, W$ and $\Lambda$. 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 $\Lambda$, 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 $X_i^{t+1}$.

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.

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 $\Lambda$, is calculated first. We use two image cues, color and edges, to calculate the observation potential as follows:

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

where $\phi_i^C$ is the observation potential for the color cue and $\phi_i^E$ is the observation potential for the edge cue. We modified the occlusion-sensitive likelihood model [10] for the observation potential with respect to the occlusion state variable $\Lambda$. 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 [10]. However, using the proposed occlusion state variable $\Lambda$, 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 term is for the visible area, and the second term is for the occluded area. The visible term is formulated as,

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

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

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

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 $i$.

We use ROM to approximate the possible range of orientation of adjacent body parts in 3D space [13] to model Kinematic potential. This kinematic relationship of the position and orientation of two adjacent body parts is formulated as Gaussian distribution with $\mu_K (= 0)$ and $\sigma_K$ to allow adjacent body parts to be loosely linked. We also learn the distribution of kinesiology from the HumanEva dataset [14].

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 $\Sigma_i$ which is a diagonal matrix with diagonal elements identical to $|\boldsymbol{\mu}_i|$.

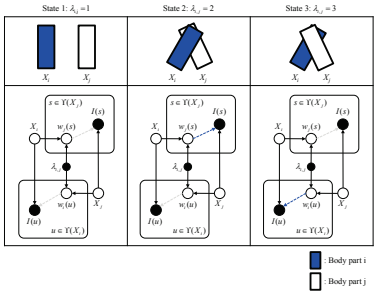



Fig. 2. Definition of three occlusion states between two body parts. The topology of $E_{O|\Lambda}$ (dashed edges in blue) changes as the value of $\lambda_{i,j}$ changes.

*2)* **Inference***:* In this paper, we use two steps to estimate the 3D body configuration and occlusion state variable separately. We can assume that $\Lambda$ 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,

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

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 $\Lambda^t$ and this increases the complexity further. In order to solve this problem, we first find overlapping (occluding) body parts with a criterion – similar to [15] – for detecting overlapping body parts as follows,

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

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,

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

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 [16]. 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 [17]. In this paper, we construct proposals for the head and torso: a face detector [18] and a head-shoulder contour detector for the torso [17].

TABLE II
THE TRACKING ERRORS IN MILLIMETERS OVER 150 FRAMES.

| Method | Walking | Jogging | Mean |
|---|---|---|---|
| PS [2] | 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 [19] | 68.67 (24.66) | 72.14 (29.62) | 69.30 (29.62) |
| our method | 88.03 (42.71) | 53.58 (17.95) | 82.83 (32.16) |

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.

## II. EXPERIMENTAL RESULTS AND ANALYSIS

We use Walking and Jogging motions of the HumanEva-I dataset [14] for evaluation. In order to evaluate the performance of occlusion state estimation, we made the ground truth data of the occlusion states for test motions. On average, manually specifying the occlusion states takes three minutes per image.

We compare four methods: (1) Pictorial Structure (PS) [2], (2) Self-occlusion Reasoning (SR) [10], (3) Mixture of Factor Analyzers (FMA) [19], and (4) the proposed method. PS and SR are chosen for comparison in terms of self-occlusion. Briefly, PS doesn't design a self-occlusion in the model while SR and the proposed method do. However, SR assumes known and fixed depth order for a target motion and, on the other hand, our method adaptively estimate occlusion states. FMA tracks 3D human pose on a manifold space with multi-view informations (camera C1-C3) while PS, SR, and the proposed method use only single view (camera C1). FMA is chosen as the state-of-the-art method. During the experiment, initializations were done manually for PS, SR, and the proposed method.

In Table II, the mean and the standard deviation of tracker 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 [19]. Since it is impossible to estimate an invisible part from a single view image, we don't count the error from completely occluded parts (for PS, SR, and the proposed method). On an average, taking invisible part into the calculation gives around 10mm higher mean error. Overall, the proposed method outperforms PS and SR for the whole motions. Since FMA tracks human pose from multi-view informations, it can exploits more useful informations such as appearance cues under self-occlusion (it is possible to roughly say that there's no self-occlusion for multi-view input since a part occluded in one camera can be seen at the other cameras), on the other hand PS, SR, and the proposed method track from a single view. Thus, FMA shows a slightly better performance.

In Table III, we analyze the complexity of 2 test motions in terms of occlusion state change (blue colored cell). The analysis is done for two part groups: 1) whole 15 body parts to measure an overall complexity and 2) 4 limbs for measuring a local complexity. We use a mean of FOC (Frame interval per Occlusion state Change) as an unit to represent how often the

TABLE III
MOTION COMPLEXITY ANALYSIS OF HUMANEVA-I (SUBJECT S2 AND CAMERA C1) AND PERFORMANCE OF OCCLUSION STATE ESTIMATION.

| Motion | Criterion | Limb | | | |
|---|---|---|---|---|---|
| | | L-ARM | R-ARM | L-LEG | R-LEG |
| Walking | Mean FOC | 5.71 | 6.35 | 6.46 | 6.31 |
| | Mean error (%) | 15.08 | 10.26 | 8.84 | 10.56 |
| Jogging | Mean FOC | 2.93 | 3.18 | 5.73 | 5.96 |
| | Mean error (%) | 17.38 | 17.38 | 8.07 | 11.44 |

occlusion state changes. According to this unit, globally and also locally, Jogging is the most complex motion in the test dataset.

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

$$E^t_{ose} = \frac{\sum_{i=1}^{K} Diff(\Lambda^t_i, \hat{\Lambda}^t_i)}{K} \quad (13)$$

where $\Lambda^t_i$ 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. (13) only upper triangular part of $\Lambda$ (and $\hat{\Lambda}$) is considered because $\Lambda$ is symmetric (strictly it is not symmetric, but conceptually it is because $\lambda_{i,j}$ and $\lambda_{j,i}$ have the same meaning). $Diff(a, b)$ 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 $\Lambda$. $K = n \times (L - 1) - n$ 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 III, the proposed method shows a 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 occlusion states of complex motion such as Walking and Jogging. Especially, it shows even better tracking performance then FMA [19] which uses multi-view informations for Jogging motion (See Table II).

## III. CONCLUSION

In this paper, we proposed a novel self-occlusion robust 3D human pose tracking method. The proposed method can track with no assumption of a known and fixed depth order about target motion. The experimental results on two motions – Walking and Jogging which have frequently changing occlusion states – shows that our method has advantages not only for tracking but also for estimating occlusion states. Particularly, the proposed method outperforms three competing methods for Jogging motion. For a future research, we will extend the proposed method for tracking human pose from interacting people and investigate robust observation model which can capture accurate information under occlusions.

## 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).

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] M. Ahmad and S.-W. Lee, "Human action recognition using shape and clg-motion flow from multi-view image sequences," *Pattern Recognition*, vol. 41, no. 7, pp. 2231–2252, 2008.

[9] A. Gupta, A. Mittal, and L. S. Davis, "Constraint integration for efficient multiview pose estimation with self-occlusions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 493–506, 2008.

[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] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.