MARKERLESS MONOCULAR TRACKING OF ARTICULATED HUMAN MOTION

*Haiying Liu**

ObjectVideo, Inc.
11600 Sunrise Valley Dr.
Reston, VA 20191

Rama Chellappa†

University of Maryland
Center for Automation Research
College Park, MD 20742

ABSTRACT

This paper presents a method for tracking general 3D general articulated human motion using a single camera with unknown calibration data. No markers, special clothes, or devices are assumed to be attached to the subject. In addition, both the camera and the subject are allowed to move freely, so that long-term view-independent human motion tracking and recognition are possible. We exploit the fact that the anatomical structure of the human body can be approximated by an articulated blob model. The optical flow under scaled orthographic projection is used to relate the spatial-temporal intensity change of the image sequence to the human motion parameters. These motion parameters are obtained by solving a set of linear equations to achieve global optimization. The correctness and robustness of the proposed method are demonstrated using Tai Chi sequences.

Index Terms— machine vision, tracking, kinematics

1. INTRODUCTION

3D human motion tracking is a very important but challenging task in many applications such as biomechanics, human interface design, surveillance systems, virtual reality. Research in this area has mainly focused on tracking relatively simple repetitive human actions such as walking or jogging, using various levels of abstraction including edges [1], contours [2], silhouettes [3, 4], texture [5], joints [6], sticks [7], blobs [8], voxel [9], depth [10], multi-view [11], and motion [12]. Most of these methods usually require more than one sequence (camera) in order to produce reasonable tracking. Methods using motion as the level of abstraction do not have such a limitation. [12] extended the method from 2D to 3D by using a twist motion model and exponential maps. A few good surveys can be found in [13, 14, 15, 16, 17, 18].

In its derivation, [12] (equation (12)) directly assumed that

$$e^{\hat{\xi}(t)_{4\times4}+\hat{\xi}'_{4\times4}} = e^{\hat{\xi}'_{4\times4}} \cdot e^{\hat{\xi}(t)_{4\times4}} \quad (1)$$

which does not hold unless the matrices $\hat{\xi}(t)$ and $\hat{\xi}'$ commute, i.e. $\hat{\xi}(t)\hat{\xi}' - \hat{\xi}'\hat{\xi}(t) = 0$. Here $\hat{\xi}(t)$ is a 4×4 matrix of twist parameters of the body pose at time t and $\hat{\xi}'$ is a 4×4 matrix related to body motion (translation and rotation). In general, $\hat{\xi}$ and $\hat{\xi}'$ do not commute. In our implementation of this approach, we found that when a twist motion model is used to track the body, the rotation and translation errors mutually propagate, i.e. the rotation error contaminates the translation estimation and vice versa. This is not surprising since the twist motion model couples the translation and rotation components, and theoretically equation (1) yields an error of $e^{\hat{\xi}'_{4\times4}} \cdot e^{\hat{\xi}(t)_{4\times4}} - e^{\hat{\xi}(t)_{4\times4}+\hat{\xi}'_{4\times4}}$.

In this paper, we propose a new solution to this problem. We have found that the twist motion model is not preferable (and not necessary) for estimating the human body base (torso) posture using motion as a level of abstraction (though it is still useful to estimate the joint motion). In our method, the motion of the torso is derived using a traditional motion model, which is expressed as a rotation $\mathbf{R}$ followed by a translation $\vec{t}$. Since the rotation $\mathbf{R}$ and the translation $\vec{t}$ are decoupled, their errors do not directly affect each other. Another contribution of this paper is that we propose a *re-initialization* scheme so that we are able to extend the degrees of freedom of each joint to be closer to those of human joints (up to 3 rather than only 1), and thus our method can track more general human motion (other than walking). By projecting the end-effector velocities to the image plane using the scaled orthographic projection model, a global linear system is constructed through the optical flow (implicitly). The least squares solution of the system yields a globally optimal estimate of the articulated body motion. The effectiveness and robustness of our algorithm are demonstrated using Tai Chi sequences, which contain general articulated human motion.

2. ARTICULATED HUMAN BODY MODEL

A general human body is modeled as a group of ellipsoids of different sizes connected by joints. This type of object is called an *articulated object*. The body parts include torso, head, upper arms, forearms, hands, thighs, and calves. The torso is regarded as the base of the articulated model. Since

*The author performed the work while at University of Maryland at College Park

†Partially funded by a grant from the office of Naval Research N00014-01.

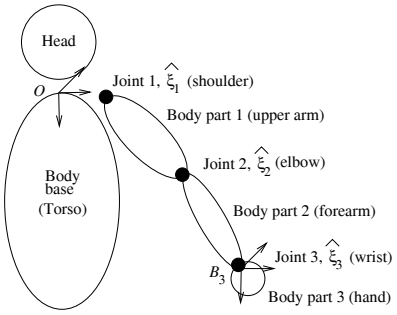

Fig. 1. Articulated human body (part) as a kinematic chain.

our goal is to track more general human motion than walking or jogging, the degrees of freedom of each joint are designed to resemble those of a real human body.

3. FORWARD KINEMATIC CHAIN

All joints of an articulated human body are spherical joints with one to three degree of freedom. In this section, we first discuss the forward kinematic chain model for an articulated body connected by pure revolute joints. The result is then expanded to spherical joints for the articulated motion tracking problem. Chasles' theorem [19] describes a systematic way of realizing any rigid body motion. Such an implementation is called a *screw motion*. For the articulated rigid human body model, the motion of a point in a body part is determined by the *forward kinematics* of the articulated human body. Assume that the body part i in its coordinate frame B_i is connected to the body base in the coordinate frame O through joints $1, 2, \cdots, i$; and the relative configurations of the pairs of adjoining parts are expressed by $\vec{\xi}_1, \vec{\xi}_2, \cdots, \vec{\xi}_i$ respectively (e.g. Figure 1). Here $\vec{\xi} = (\vec{v}, \vec{\omega})$ is the screw motion vector, with the rotation axis defined by $\vec{\omega}$ and the translation related to $\vec{v}$. The rigid motion of the body part i in the frame O is

$$\mathbf{M}_{ob_i}(\theta) = e^{\hat{\xi}_1\theta_1} \cdot e^{\hat{\xi}_2\theta_2} \cdot \cdots \cdot e^{\hat{\xi}_i\theta_i} \cdot \mathbf{M}_{ob_i}(0), \qquad (2)$$

where $\hat{\xi} = \begin{bmatrix} \hat{\omega} & \vec{v} \\ 0 & 0 \end{bmatrix} \in \mathbf{x}R^{4\times 4}$. Here, $\mathbf{M}_{ob_i}$ is the rigid body transformation between the frame B_i and the frame O when the body parts are in the reference configuration with $\theta = 0$. Figure 1 is an example of such a kinematic chain. The rigid motion of the hand in the frame O can be conveniently calculated by $\mathbf{M}_{ob_3}(\theta) = e^{\hat{\xi}_1\theta_1} \cdot e^{\hat{\xi}_2\theta_2} \cdot e^{\hat{\xi}_3\theta_3} \cdot \mathbf{M}_{ob_3}(0)$, if all joints are modeled as revolute joints. Equation (2) is called the *product of exponentials formula*. It can be proved [19] that $\mathbf{M}_{ob_i}$ is independent of the order in which the joints are rotated. In a more general (natural) model, the shoulder has three degrees of freedom, and the elbow and wrist have two degrees of freedom each. Physically, this can be constructed by combining two or three revolute joints with their axes orthogonal to each other and all intersecting at one point. Thus the combined motion is

$$\mathbf{M}_{bi} = e^{\hat{\xi}_{i_1}\theta_{i_1}} \cdot e^{\hat{\xi}_{i_2}\theta_{i_2}} \cdot e^{\hat{\xi}_{i_3}\theta_{i_3}}. \qquad (3)$$

In Figure 1, the rigid motion of the hand in frame O is in the form $\mathbf{M}_{ob_3}(\theta) = \left(e^{\hat{\xi}_{11}\theta_{11}} e^{\hat{\xi}_{12}\theta_{12}} e^{\hat{\xi}_{13}\theta_{13}}\right)_{shoulder} \cdot \left(e^{\hat{\xi}_{21}\theta_{21}} e^{\hat{\xi}_{22}\theta_{22}}\right)_{elbow} \cdot \left(e^{\hat{\xi}_{31}\theta_{31}} e^{\hat{\xi}_{32}\theta_{32}}\right)_{wrist} \cdot \mathbf{M}_{ob_3}(0)$. The motions of other body parts can be computed similarly.

4. END EFFECTOR VELOCITY OF THE KINEMATIC CHAIN

In a kinematic chain, joint velocities

$$\dot{\vec{\theta}} = [\dot{\theta}_1, \dot{\theta}_2, \cdots, \dot{\theta}_i]^T \qquad (4)$$

are mapped to end-effector velocities $\vec{V}^o_{ob_i}$ by

$$\vec{V}^o_{ob_i} = \mathbf{J}^o_{ob_i}(\vec{\theta})\dot{\vec{\theta}} \qquad (5)$$

where $\mathbf{J}^o_{ob_i}(\theta)$ is the *spatial manipulator Jacobian* $\mathbf{J}^o_{ob_i}(\vec{\theta})$: $\mathbf{x}R^n \to \mathbf{x}R^{6\times n}$. It is defined as follows [19]:

$$\mathbf{J}^o_{ob_i}(\vec{\theta}) = \left[\vec{\xi}_1, \vec{\xi}'_2, \cdots, \vec{\xi}'_i\right], \vec{\xi}'_i = \mathrm{Ad}_{e^{\hat{\xi}_1\theta_1} \cdots\cdots e^{\hat{\xi}_{i-1}\theta_{i-1}}} \vec{\xi}_i. \qquad (6)$$

Here $\vec{\theta} = [\theta_1, \theta_2, \cdots, \theta_i]^T$ is the joint angle configuration, $\vec{\xi}_i$ is the ith joint axis orientation configuration, and $\mathrm{Ad}_{\mathbf{M}}$ is the *adjoint transformation* of the rigid motion $\mathbf{M}$: $\mathrm{Ad}_{\mathbf{M}} = \begin{bmatrix} \mathbf{R} & \hat{t}\mathbf{R} \\ \mathbf{0}_{3\times 3} & \mathbf{R} \end{bmatrix}$ where $\mathbf{M}$ is in the form of $\mathbf{M} = \begin{bmatrix} \mathbf{R} & \vec{t} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}$. $\mathbf{M}$ defines a rigid motion as a rotation $\mathbf{R}$ followed by a translation $\vec{t}$.

5. ARTICULATED BODY TRACKING USING OPTICAL FLOW

5.1. Tracking the body base

In this sub-section, we focus on deriving the equations for tracking a single body base (torso). If the body base can be successfully tracked, all the other body parts can be tracked through the kinematic chain given in Section 3.

Notice that the only input data is a video sequence and the motion is observed through the optical flow. An appropriate projection model is needed to project the 3-D motion into the 2-D optical flow. When the camera calibration parameters are unknown and the subject is not too close to the camera, the scaled orthographic projection is a good approximation. Denote by O the camera coordinate frame and by B the body base (part) coordinate frame. A point $\vec{p}_b = [X_b, Y_b, Z_b]^T$ in frame B corresponds to $\vec{p}_o = [X_o, Y_o, Z_o]^T$ in frame O. Projecting the point onto the image plane, its correspondent image coordinates $\vec{p}_{im} = [x_{im}, y_{im}]^T$ are

$$\vec{p}_{im} = \begin{bmatrix} x_{im} \\ y_{im} \end{bmatrix} = s \cdot [\mathbf{I}_{2\times 2} \mid \mathbf{0}_{2\times 2}] \cdot \vec{p}_o = s \cdot \mathbf{P} \cdot \vec{p}_o \qquad (7)$$

where s is the scale and

$$\mathbf{P} = [\mathbf{I}_{2\times 2} \mid \mathbf{0}_{2\times 2}] \tag{8}$$

is the orthographic projection matrix.

Assume that from time t to time $t+1$, the point $\vec{p}_o$ rotates by $e^{\Delta\hat{\omega}}$ and then translates by $\Delta\vec{v}$, and the scale is changed by $\Delta s = s's(t)$. Note that for small $\Delta\vec{\omega}$,

$$\begin{aligned} e^{\Delta\hat{\omega}} &= \mathbf{I} + \Delta\hat{\omega} + \frac{\Delta\hat{\omega}^2}{2} + \cdots + \frac{\Delta\hat{\omega}^n}{n!} + \cdots \\ &\approx \mathbf{I} + \Delta\hat{\omega} \end{aligned} \tag{9}$$

Under scaled orthographic projection, the optical flow $\vec{u} = [u_x, u_y]^T$ of its corresponding point in the image plane is

$$\begin{aligned} \begin{bmatrix} u_x \\ u_y \end{bmatrix} &= \begin{bmatrix} x_{im}(t+1) - x_{im}(t) \\ y_{im}(t+1) - y_{im}(t) \end{bmatrix} \\ = s(t)\cdot\mathbf{P}\cdot((s'\mathbf{I} + (1+s')\Delta\hat{\omega})\,\vec{p}_o + (1+s')\Delta\vec{v}) \end{aligned} \tag{10}$$

Let $\hat{\omega}' = (1+s')\Delta\hat{\omega}, \vec{v}' = (1+s')\Delta\vec{v}$, where $\hat{\omega}'$ and $\vec{v}'$ are in the form $\hat{\omega}' = [\omega'_x, \omega'_y, \omega'_z]^T, \vec{v}' = [v'_x, v'_y, v'_z]^T$. Equation (10) becomes

$$\begin{aligned} \begin{bmatrix} u_x \\ u_y \end{bmatrix} &= s(t)\cdot\mathbf{P}\cdot((s'\mathbf{I} + \hat{\omega}')\,\vec{p}_o + \vec{v}') \\ = s(t)\left(\begin{bmatrix} s' & -\omega'_z & \omega'_y \\ \omega'_z & s' & -\omega'_x \end{bmatrix}\vec{p}_o + \begin{bmatrix} v'_x \\ v'_y \end{bmatrix}\right) \end{aligned} \tag{11}$$

Recall that $\vec{p}_o = [X_o, Y_o, Z_o]^T$. Substituting the flow $\vec{v} = [u_x, u_y]^T$ in the optical flow constraint by (11), we get

$$(I_x, I_y)\cdot\vec{v} + I_t = 0 \tag{12}$$

$$\begin{aligned} \Rightarrow s(t)\left((I_xX_o + I_yY_o)s' + I_xv'_x + I_yv'_y - \right. \\ \left. I_yZ_o\omega'_x + I_xZ_o\omega'_y + (I_yX_o - I_xY_o)\omega'_z\right) = -I_t \end{aligned} \tag{13}$$

For N pixels on the body base, we have the linear system

$$s(t)\cdot\mathbf{A}\cdot\vec{\delta} = -\vec{I}_t \tag{14}$$

where

$$\mathbf{A} = \left[\vec{A}_1, \vec{A}_2, \cdots, \vec{A}_N\right]^T \tag{15}$$

$$\begin{aligned} \vec{A}_i = [&I_{xi}X_{oi} + I_{yi}Y_{oi}, I_{xi}, I_{yi}, \\ &-I_{yi}Z_{oi}, I_{xi}Z_{oi}, I_{yi}X_{oi} - I_{xi}Y_{oi}]^T \end{aligned} \tag{16}$$

$$\vec{\delta} = \left[s', v'_x, v'_y, \omega'_x, \omega'_y, \omega'_z\right]^T \tag{17}$$

$$\vec{I}_t = [I_{t1}, I_{t2}, \cdots, I_{tN}]^T \tag{18}$$

The motion (a rotation $\mathbf{R}$ followed by a translation $\Delta\vec{v}$) of the body base can then be calculated through the least-squares solution of (14) by $\mathbf{R} = e^{\frac{\hat{\omega}'}{1+s'}}, \Delta\vec{v} = \frac{\vec{v}'}{1+s'}$ Note that the translation along the Z-axis cannot be detected, due to the scaled orthographic projection, but it is reflected in the change of scale s.

5.2. Tracking the other body parts

Recalling (5), (6), (7), and (8), the optical flow contributed by the scaled orthographic projection of the kinematic chain is $\begin{bmatrix} u_x \\ u_y \end{bmatrix} = s(t)\cdot\mathbf{P}\cdot\vec{V}^o_{ob_i}\cdot\vec{p}_o$ Therefore, the total optical flow of the ith body part is $\begin{bmatrix} u_x \\ u_y \end{bmatrix} = s(t)\cdot\mathbf{P}\cdot\left((s'\mathbf{I} + \hat{\omega}')\,\vec{p}_o + \vec{v}' + \vec{V}^o_{ob_i}\cdot\vec{p}_o\right)$ Substituting it into equation (12), we get

$$s(t)\cdot\left(\vec{A}_i^{\,T}\cdot\vec{\delta} + \vec{B}_i^{\,T}\cdot\dot{\vec{\theta}}\right) = -\vec{I}_t \tag{19}$$

where $\vec{A}_i$ is defined in (16), $\vec{\delta}$ is defined in (17), $\vec{I}_t$ is defined in (18), $\dot{\vec{\theta}}$ is defined in (4) for all K joints in the articulated body model, and $\vec{B}_i$ is defined as $\vec{B}_i = [B_{i1}, B_{i2}, \cdots, B_{iK}]^T$. Here, $B_{ij} = [I_x, I_y]\cdot\mathbf{P}\cdot\hat{\xi}'_j\cdot\vec{p}_o$ when joint ξ_j is on the kinematic chain that affects pixel i, 0 otherwise.

Let

$$\vec{\Theta} = \left[s', v'_x, v'_y, \omega'_x, \omega'_y, \omega'_z, \dot{\theta}_1, \dot{\theta}_2, \cdots, \dot{\theta}_K\right]^T. \tag{20}$$

For all N pixels on the ith body part, (19) can be re-written as

$$s(t)\cdot[\mathbf{A}_i|\mathbf{B}_i]\cdot\vec{\Theta} = -\vec{I}_t \tag{21}$$

where $\mathbf{A}_i$ is defined in (15), and $\mathbf{B}_i = \left[\vec{B}_{i,1}, \vec{B}_{i,2}, \cdots, \vec{B}_{i,N}\right]^T$ for the ith body part.

5.3. Tracking the whole body

Stacking (21) for all $i = 1, 2, \cdots, M$ body parts together, we get the global linear system

$$s(t)\cdot[\mathbf{A}|\mathbf{B}]\cdot\vec{\Theta} = -\vec{I}_t \tag{22}$$

where $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_M]^T, \mathbf{B} = [\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_M]^T$. The least squares solution of (22) yields the complete body motion.

6. EXPERIMENTS

The ellipsoid model was used in our experiments. We tested our method on Tai Chi sequences. The original sequence was encoded in low-quality mpeg format. Both the background and the Tai Chi performer were moving. The loose and relatively un-textured clothes made tracking even more challenging. The body was defined as a 28 degree-of-freedom (DOF) kinematic structure. During the Tai Chi performance, the brightness intensity of the torso and limbs often changed due to the severe shadow. The initial pose of the body model was pre-determined. The results show that the high-DOF human body model, performing complex actions, was tracked robustly. The results for two sequences are shown in Figures 2 and 3. It can be observed that the 3D rotations of the torso, arms, and legs are well tracked. The result shows the tracking is reasonably good.

7. CONCLUSION

We have proposed a closed-form solution for the markerless monocular articulated motion tracking problem. Optical flow is exploited implicitly to use all pixels of the body parts as dense natural built-in markers. We have shown that the twist motion model introduces errors and thus is not preferable (and not necessary) to estimate the motion of the articulated human body base (torso). We have given a new derivation using motion as a level of abstraction. Together with the exponential map and kinematic chain, this provides constraints that facilitate robust tracking. The degrees of freedom of all joints are designed to resemble those of a real human body, so that human actions more general than walking and jogging can be tracked. A re-initialization scheme is also proposed to handle the singularities of exponential coordinates. Without explicitly using any features, this allows both the objects and the camera to move freely. It also avoids excessive pre-processing (such as background subtraction, edge/contour/silhouette detection, etc.), which are usually not stable or even practically feasible. Limited experiments show the effectiveness and robustness of our method.

8. REFERENCES

[1] D.M. Gavrila and L.S. Davis, "3-D model-based tracking of humans in action: A multi-view approach," *CVPR*, pp. 73–80, 1996.

[2] D. Meyer, J. Denzler, and H. Niemann, "Model based extraction of articulated ojects in image sequences for gait analysis," *ICIP*, pp. 78–81, 1997.

[3] Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with silhouettes," *ICCV*, pp. 716–721, 1999.

[4] K.M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *IEEE CVPR*, June 2003.

[5] H. Sidenbladh, F. De la Torre, and M.J. Black, "A framework for modeling the appearance of 3D articulated figures," *Intl. Conf. on Automatic Face and Gesture Recognition*, pp. 368–375, 2000.

[6] O. Munkelt, C. Ridder, D. Hansel, and W. Hafner, "A model driven 3D image interpretation system applied to person detection in video images," *ICPR*, pp. 307–314, 1998.

[7] Y. Guo, G. Xu, and S. Tsuji, "Tracking human body motion based on a stick figure model," *Journal of Visual Communication and Image Representation*, pp. 1–9, 1994.

[8] C.R. Wren and A.P. Pentland, "Dynamic models of human motion," *Intl. Conf. on Automatic Face and Gesture Recognition*, pp. 22–27, 1998.

[9] Ivana Mikic, Mohan Trivedi, Edward Hunter, and Pamela Cosman, "Human body model acquisition and tracking using voxel data," *International Journal of Computer Vision*, vol. 53, no. 3, 2003.

[10] R. Plänkers and Pascal Fua, "Articulated soft objects for video-based body modeling," *ICCV*, pp. 394–401, 2001.

[11] Aravind Sundaresan, Amit RoyChowdhury, and Rama Chellappa, "Multiple view tracking of human motion modelled by kinematic chains," in *International Conference on Image Processing, Singapore*, 2004.

[12] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," *CVPR*, pp. 8–15, 1998.

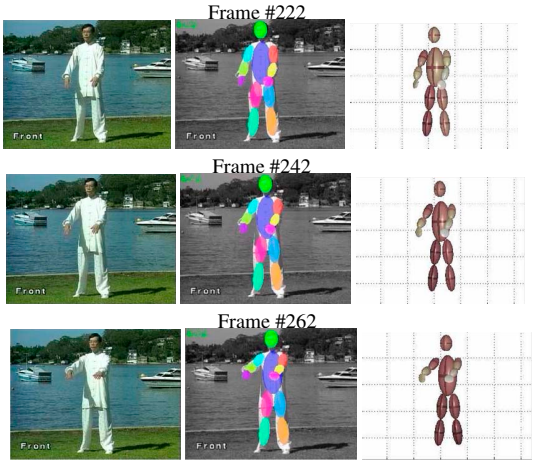

Fig. 2. Tai Chi sequence with self-occlusion and arms not parallel to the image plane.

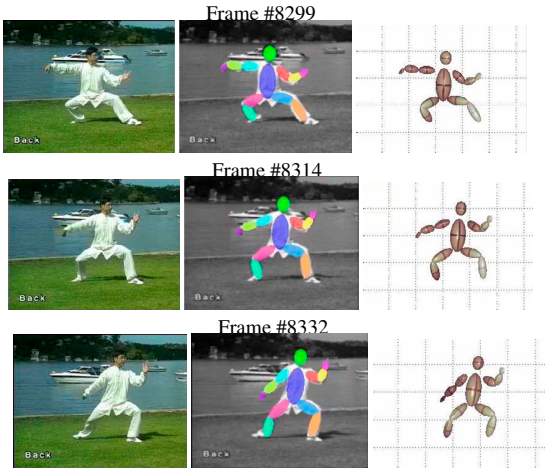

Fig. 3. Tai Chi sequence with large motion.

[13] T.B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *CVIU*, pp. 231–268, 2001.

[14] Joseph Bray, "Markerless based human motion capture: A survey," *Lab report*, 2003.

[15] T. Tan L. Wang, W. Hu, "Recent developments in human motion analysis," *Pattern Recognition*, vol. 36, no. 3, pp. 585–601, March 2003.

[16] N. Krahnstoever and R. Sharma, "Articulated models from video," in *Computer Vision and Pattern Recognition*, 2004, pp. 894–901.

[17] Xiangyang Lan and Daniel P. Huttenlocher, "A unified spatio-temporal articulated model for tracking.," in *CVPR*, 2004, pp. 722–729.

[18] Xiaofeng Ren, Alexander C. Berg, and Jitendra Malik, "Recovering human body configurations using pairwise constraints between parts," in *International Conference on Computer Vision*, 2005.

[19] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulations*, CRC Press, 1994.