Computer Vision and Image Understanding **81,** 269–284 (2001)
doi:10.1006/cviu.2000.0888, available online at http://www.idealibrary.com on IDEAL®

Estimating Anthropometry and Pose from a Single Uncalibrated Image

Carlos Barrón and Ioannis A. Kakadiaris

Department of Computer Science, University of Houston, 4800 Calhoun, Houston, Texas 77204-3475
E-mail: cbarron@uh.edu, ioannisk@uh.edu

Accepted July 29, 2000

In this paper, we present a four-step technique for simultaneously estimating a human's anthropometric measurements (up to a scale parameter) and pose from a single uncalibrated image. The user initially selects a set of image points that constitute the projection of selected landmarks. Using this information, along with a priori statistical information about the human body, a set of plausible segment length estimates is produced. In the third step, a set of plausible poses is inferred using a geometric method based on joint limit constraints. In the fourth step, pose and anthropometric measurements are obtained by minimizing an appropriate cost function subject to the associated constraints. The novelty of our approach is the use of anthropometric statistics to constrain the estimation process that allows the simultaneous estimation of both anthropometry and pose. We demonstrate the accuracy, advantages, and limitations of our method for various classes of both synthetic and real input data.

1. INTRODUCTION

Video-based three-dimensional human motion tracking is an important and challenging research problem. Its importance stems from numerous applications such as: (1) performance measurement for human factors engineering, (2) posture and gait analysis for training athletes and physically challenged persons, (3) human body, hands, and face animation, and (4) automatic annotation of human activities in video databases. The challenges toward the general applicability of a vision-based three-dimensional tracking system on real data include the following:

- *Data from one camera only*: There are several applications for which the video recordings from only one view are available (e.g., for analyzing the motion of famous artists in historical recordings). In addition, the camera might be moving, possibly zooming in and out.

1077-3142/01 $35.00

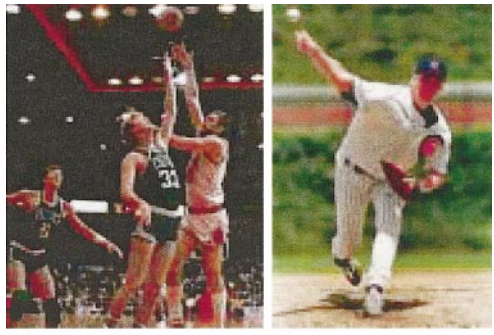

FIG. 1. (a) Instance of an image that can be handled by our algorithm. (b) Instance of an image that cannot be handled by our algorithm.

- *Model acquisition*: There is no such thing as an "average" human and that makes the selection of a geometric model for model-based tracking difficult.
- *Modeling*: The human models that are currently used for motion estimation do not incorporate statistical anthropometric information.

Our goal is to develop a model-based system for tracking humans from monocular images. In this paper, we present a technique for simultaneous anthropometry and pose estimation from the first frame of an image sequence. The input to the algorithm is the image coordinates of the visible landmarks from the human subject (as selected by the user) in the image under examination (Fig. 6a). The output is the subject's anthropometric measurements (up to a scale parameter) and his/her pose in the specific image (Fig. 6b). By the term "up to a scale," we refer to the fact that from a single uncalibrated camera we cannot infer absolute lengths (like "upper-leg-length" and "shoulder-width") but only ratios of lengths. Therefore, in the following when we refer to the estimation of the anthropometric measurements, we imply the estimation of ratios of lengths like "upper-leg-length" over "shoulder-width." The novelty of our approach is the use of anthropometric statistics to constrain the estimation process. The impact of our method lies in the ability to semi-automate the initialization phase for model-based human tracking methods from a single camera. As will be explained in later sections, our method can handle images like the one depicted in Fig. 1a, but not images like the one depicted in Fig. 1b.

The remainder of this paper describes our technique in more detail. In Section 2, we review prior work in the area, while in Section 3 we formulate the problem and we offer a detailed analysis of the geometric and statistical relationships. In Section 4, we describe our method in detail, and in Section 5 we present a number of results from our system.

2. PRIOR WORK

Two of the challenges in model-based human tracking algorithms are: (1) the acquisition of an accurate human body model that will be employed as the model, and (2) the initialization of the model in the first frame of the image sequence. Concerning model acquisition, existing approaches use either models of the human body whose parts are approximated with simple shapes and their dimensions have been manually measured [10, 20] or models whose shape and/or dimensions have been determined based on camera input data. In this

second category, methods have been developed to obtain models of human body parts from multiple cameras [11, 12, 15] or range data [8]. Concerning posture estimation and tracking, methods have been presented that use either one [4, 6, 17, 21], or multiple cameras [1, 7, 9, 13, 14, 16]. However, in most of the existing tracking approaches the user specifies an approximate position and posture for the human model at the first frame of the image sequence [6, 14, 19]. In contrast, Bregler and Malik [4], for the initialization step of their human tracking method, minimize a cost function over position, angles, and body dimensions. In particular, a user selects the 2D joint locations and then a 3D pose is found by minimizing the sum of the squared differences between the projected model joint locations and the corresponding model joint locations. The authors mention that they had good results with a quasi-Newton method and a mixed quadratic and line search procedure. However, no information is provided about the accuracy and repeatability of their method, nor for what class of postures and human body dimensions does their method succeed. The contribution of our paper is a systematic study and a technique that takes into consideration statistical anthropometric information to constrain the estimation process.

3. ANALYSIS

In this section, first we formulate the problem, then we present a stick human body model (SM) that incorporates statistical anthropometric information, and finally we provide a detailed analysis of the geometrical and statistical relationships of the SM's segments.

3.1. Problem Statement

The human musculosketelal system is composed of a series of jointed links, which can be approximated as rigid bodies. Human motion estimation is aimed at quantitatively describing the spatial motion of body segments and the movements of the joints connecting those segments. A hallmark of the individuality of people we encounter daily results from the variation of their anthropometric measurements. If we assume that we have no anthropometric information for the subject that we are observing, the problem of anthropometry and pose estimation from a single image can be formulated as follows: Given a set of points in an image that correspond to the projection of landmark points of a human subject, estimate both the anthropometric measurements (up to a scale) of the subject and his/her pose that best match the observed image.

3.2. Stick Human Body Model

For the purposes of this research, we have developed a generic stick human body model (Fig. 2) inspired by the human body model employed at the Center for Human Modeling and Simulation at University of Pennsylvania [3]. The model consists of a set of segments connected by joints. Specifically, a stick model is a tree $(s, \mathcal{S}, \mathcal{A})$, where $\mathcal{S}$ is a set of sites/landmarks and $\mathcal{A}$ is a collection of edges (segments) with endpoints in $\mathcal{S}$, and $s \in \mathcal{S}$ is the root. In our case, $\mathcal{A} =$ {HD, RY, LY, NK, UT, RC, LC, RUA, LUA, RLA, LLA, RHD, LHD, LT, RHP, LHP, RUL, LUL, RLL, LLL, RF, LF} as enumerated in Table 1, and the set of landmarks consists of a set of joints $\mathcal{J} =$ {at, sp, la, lc, le, lh, lk, ls, lw, ra, rc, re, rh, rk, rs, rw, wt} (information about the SM's joints is provided in Table 2) and other landmarks $\mathcal{M} =$ {ry (right eye), ly (left eye), rhd (base of the right middle finger),

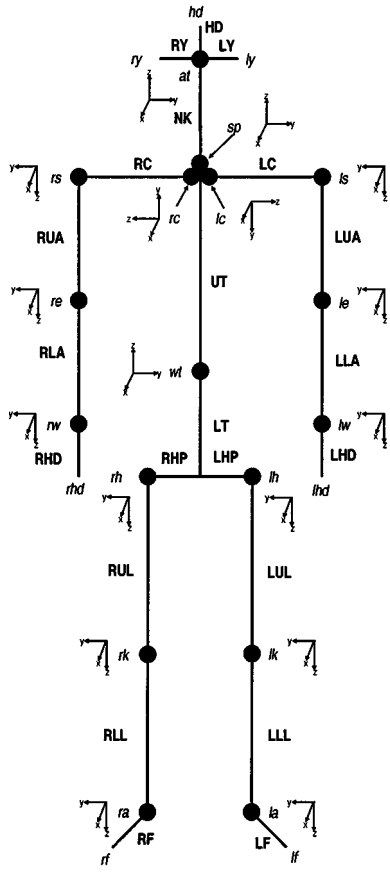

FIG. 2. Stick human body model (SM) and its associated coordinate systems.

lhd (base of the left middle finger), rf (tip of the right foot), if (tip of the left foot)} ($\mathcal{S} = \mathcal{J} \cup \mathcal{M}$).

A local coordinate system is attached to each body part. The kinematics are represented by a transformation tree whose root is the subject's coordinate system and whose leaves are the coordinate systems of head, hands, and feet. The origin of the subject's coordinate system is the waist joint. Figure 2 depicts the local coordinate systems of the stick human model, which corresponds to the joints listed in Table 2. Note that every joint has translational and rotational degrees of freedom. The joint's translational degrees of freedom allow for segment scaling, and they are restricted by anthropometric proportionality constraints as explained in Section 4.3. Each rotational degree of freedom has an upper limit and a lower limit that restricts the pose estimation to plausible human postures. The default data for the joints are extracted from [18].

3.3. Geometric Relationships

In this section, we will examine the foreshortening of the body segments in the image, under the assumption of scaled orthographic projection. Let $\mathbf{c} = [X_c, Y_c, Z_c]^\top$ be the origin of the camera (see Fig. 3) and let's assume that the image plane is located at Z_{im} along the Z axis of the camera. As known, under scaled orthographic projection the point $\mathbf{P}_1 = [X_1, Y_1, Z_1]^\top$

TABLE 1
Names of the SM's Segments

ID	Segment	ID	Segment
HD	Head	NK	Neck
LY	Left eye	RY	Right eye
LT	Lower torso	UT	Upper torso
LC	Left clavicle	RC	Right clavicle
LUA	Left upper arm	RUA	Right upper arm
LLA	Left lower arm	RLA	Right lower arm
LHD	Left hand	RHD	Right hand
LHP	Left hip	RHP	Right hip
LUL	Left upper leg	RUL	Right upper leg
LLL	Left lower leg	RLL	Right lower leg
LF	Left foot	RF	Right foot

(see Fig. 3) projects to the point $\mathbf{p}_1 = [x_1, y_1, z_1]^\top = [X_c + \lambda_1(X_1 - X_c), Y_c + \lambda_1(Y_1 - Y_c), Z_{\text{im}}]^\top$, where $\lambda_1 = (Z_{\text{im}} - Z_c)/(Z_1 - Z_c)$. Similarly, the point $\mathbf{P}_2 = [X_2, Y_2, Z_2]^\top$ projects to the point $\mathbf{p}_2 = [x_2, y_2, z_2]^\top = [X_c + \lambda_2(X_2 - X_c), Y_c + \lambda_2(Y_2 - Y_c), Z_{\text{im}}]^\top$, where $\lambda_2 = (Z_{\text{im}} - Z_c)/(Z_2 - Z_c)$. If we assume that this point lies on the same plane (normal to the camera Z axis) as the point $\mathbf{P}_1$, then $\lambda_1 = \lambda_2$. Thus, for any point on the line $\overline{\mathbf{P}_1\mathbf{P}_2}$, its projection is given by the equation $[x, y]^\top = \lambda_1 \mathbf{S}[X, Y, Z]^\top$, where

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Similarly, for any point on the line $\overline{\mathbf{P}_3\mathbf{P}_4}$, its projection is given by the equation $[x, y]^\top = \lambda_3 \mathbf{S}[X, Y, Z]^\top$, where $\lambda_3 = (Z_{\text{im}} - Z_c)/(Z_3 - Z_c) = (Z_{\text{im}} - Z_c)/(Z_4 - Z_c) = \lambda_4$. If α_z is

TABLE 2
Information Related to the Joints of the Stick Model

ID	Joint	From	To	DOF	PR
at	atlanto occipital	NK	HD	Tz*Rz*Ry*Rx	3
sp	solar plexus	UT	NK	Tz*Ry*Rz*x	2
la	left ankle	LLL	LF	Tx*Rz*Rx*Ry	4
lc	left clavicle	UT	LC	Tz*Rx*Ry	3
le	left elbow	LUA	LLA	Tz*Ry	5
lh	left hip	LT	LUL	Tz*Rz*Rx*Ry	2
lk	left knee	LUL	LLL	Tz*R-y	3
ls	left shoulder	LC	LUA	Tz*Rz*Rx*Ry	4
lw	left wrist	LLA	LHD	Tz*Ry*Rx*Rz	6
ra	right ankle	RLL	RF	Tx*R-z*R-x*Ry	4
rc	right clavicle	UT	RC	Tz*R-x*Ry	3
re	right elbow	RUA	RLA	Tz*Ry	5
rh	right hip	LT	RUL	Tz*R-z*R-x*Ry	2
rk	right knee	RUL	RLL	Tz*R-y	3
rs	right shoulder	RC	RUA	Tz*R-z*R-x*Ry	4
rw	right wrist	RLA	RHD	Tz*Ry*R-x*R-z	6
wt	waist	LT	UT	Tz*Ry*Rz*Rx	1

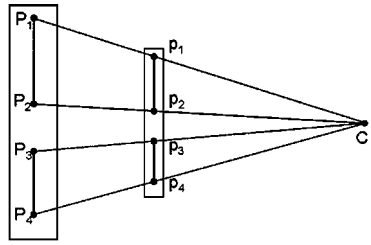

FIG. 3. Notation pertaining to Proposition 3.1.

a real number such that

$$(1+\alpha_z) = \frac{Z_1 - Z_c}{Z_3 - Z_c}, \tag{1}$$

then $Z_3 - Z_c = (Z_1 - Z_c)/(1+\alpha_z)$ and $\lambda_3 = \lambda_1(1+\alpha_z)$. Therefore, the scaled orthographic projection for the points of $\overline{\mathbf{P}_3\mathbf{P}_4}$ is given by $[x, y]^\top = \lambda_1(1+\alpha_z)\mathbf{S}[X, Y, Z]^\top$. Let $L_{12} = \|\mathbf{P}_2 - \mathbf{P}_1\|$ and $l_{12} = \|\mathbf{p}_2 - \mathbf{p}_1\|$. Then

$$l_{12} = ((x_2 - x_1)^2 + (y_2 - y_1)^2)^{\frac{1}{2}} = \lambda_1((X_2 - X_1)^2 + (Y_2 - Y_1)^2)^{\frac{1}{2}} = \lambda_1 L_{12}.$$

Similarly, we can obtain that $l_{34} = \lambda_3 L_{34}$, where $L_{34} = \|\mathbf{P}_4 - \mathbf{P}_3\|$ and $l_{34} = \|\mathbf{p}_4 - \mathbf{p}_3\|$. Using the relation $\lambda_3 = \lambda_1(1+\alpha_z)$, we obtain that $l_{34} = \lambda_1(1+\alpha_z)L_{34}$. Finally, the ratio of l_{12} and l_{34} is given by

$$\frac{l_{12}}{l_{34}} = \frac{\lambda_1 L_{12}}{\lambda_1(1+\alpha_z)L_{34}} = \frac{L_{12}}{(1+\alpha_z)L_{34}},$$

which implies the relation

$$\frac{L_{12}}{L_{34}} = (1+\alpha_z)\frac{l_{12}}{l_{34}}, \tag{2}$$

which suggests the following proposition.

PROPOSITION 3.1. *For segments that lie in planes almost parallel to the image plane, the ratio of segment lengths in 3D is similar to the ratio of the lengths of the corresponding segments projected to the image plane.*

Proof. Since the segments lie in planes almost parallel to the image plane, α_z is very small. Thus, the result is obtained from Eq. (2). ■

3.4. Building a Cadre Family

Using the anthropometric measurements in [18], we build for our SM a cadre family, also known as a boundary family [2]. The cadre family is a multivariate representation of the extremes of the population distribution. It has the ability to span the multivariate space in a systematic fashion and to capture a significant amount of the variance in the space using a small number of sample human models. The probability density function of the multivariate normal distribution is defined by

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{s})^\top \Sigma^{-1}(\mathbf{x} - \mathbf{s})\right], \tag{3}$$

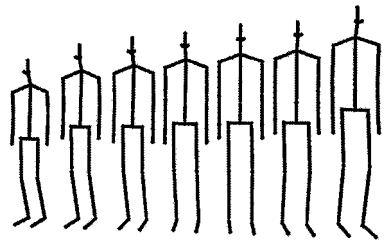

FIG. 4. Sample models from the distribution of SMs (7 out of 143).

where k is the number of dimensions. In our case, the variables are the lengths of the 22 segments of our stick model (see Table 1), $\mathbf{x}$ is a random vector, and $\mathbf{s}$ and $\mathbf{\Sigma}$ are the mean and the covariance matrices of the population, respectively.

The quadratic form $Q(\mathbf{x}) = (\mathbf{x} - \mathbf{s})^\top \mathbf{\Sigma}(\mathbf{x} - \mathbf{s})$ defines a hyperellispoid surface around $\mathbf{s}$, whose shape depends on $\mathbf{\Sigma}$. We compute the principal components of $\mathbf{\Sigma}$, and we select seven $\mathbf{E}_i$ $(i = 1, \ldots, 7)$ eigenvectors with the largest eigenvalues such that $\lambda_1 > \lambda_2 > \cdots > \lambda_7$. Note that each eigenvector corresponds to the original 22 variables associated with the lengths as $\mathbf{E}_i = [e_{i1}, e_{i2}, \ldots, e_{i22}]^\top$. In addition, all linear combinations $\sum_{i=1}^{7} \alpha_i \mathbf{E}_i + \mathbf{s}$ constrained by the relation

$$\sum_{i=1}^{7} \alpha_i^2 \leq 1 \tag{4}$$

lie in the interior of the hyperellispoid related with $\mathbf{\Sigma}$ and $\mathbf{s}$. By selecting the positive nonzero coefficients that satisfy Eq. (4) ($|\alpha_1| = |\alpha_2| = \cdots = |\alpha_7| \neq 0$), we can build a family of stick models that correspond to all equidistant discrete combinations of the selected eigenvectors: $\sum_{i=1}^{7} \beta_i \mathbf{E}_i + \mathbf{s}$, where $\beta_i \in \{-\frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}\}$. Furthermore, we add in our cadre family $\mathbf{s}$ and the axial points $\pm\mathbf{E}_i + \mathbf{s}$, $i = 1, \ldots, 7$. The total number of SMs produced is 143 (in general this procedure gives $2^n + 2n + 1$ components, where n is the number of principal component vectors kept). A sample of these SMs is depicted in Fig. 4.

3.5. Determining a Covering Set

In this section, we describe our algorithm for determining the set of anthropometric proportions that will be used by our iterative estimator to reach an anthropometrically plausible solution. In order to reduce the number of variables, we assume that the left and right sides of the human body are symmetrical, and we only consider the left side. Also, we do not employ the segments associated with the eyes, the hands, and the atlanto occipital joint. Thus, we focus our statistical modeling on eight segments of the SM as enumerated in Table 3. In the following, let $\mathcal{L} = \{l_i\}_{i=1}^{8}$ be the set of segment lengths, and $\mathcal{R} = \{r_k\}_{k=1}^{28}$ be the set of the corresponding ratios of segment lengths. These ratios are computed by dividing the lengths of any two different segments that belong to $\mathcal{L}$.

TABLE 3
The Segments Used for Computing the Covering Set

l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8
UT + LT	LC	LUA	LLA	LHP	LUL	LLL	LF

TABLE 4
Segment Lengths and Ratios Associated with Our Cadre Family of SMs

$l_{1,1}$	$l_{2,1}$	...	$l_{8,1}$	$\Rightarrow$	$r_{1,1}$	$r_{2,1}$	...	$r_{28,1}$
$l_{1,2}$	$l_{2,2}$	...	$l_{8,2}$	$\Rightarrow$	$r_{1,2}$	$r_{2,2}$	...	$r_{28,2}$
$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
$l_{1,143}$	$l_{2,143}$	...	$l_{8,143}$	$\Rightarrow$	$r_{1,143}$	$r_{2,143}$	...	$r_{28,143}$
	Mean			$\Rightarrow$	$\mu(r_1)$	$\mu(r_2)$	...	$\mu(r_{28})$
	Variance			$\Rightarrow$	$\sigma(r_1)$	$\sigma(r_2)$	...	$\sigma(r_{28})$

We form these $\binom{8}{2} = 28$ ratios as follows:

$$r_{k,q} = \begin{cases} \frac{l_{m,q}}{l_{n,q}} & \text{if } \mu(l_n) > \mu(l_m) \\ \frac{l_{n,q}}{l_{m,q}} & \text{otherwise,} \end{cases}$$

where $1 \leq m < n \leq 8$, $k = 1, \ldots, 28$, and $q = 1, \ldots, 143$. The index formula that relates k with m and n is given by: $k = n - m + (m-1)(8 - \frac{m}{2})$.

In Table 4 the rows depict that the lengths of the eight selected segments will be used to produce the SM's ratios, and the associated means and variances. The segment lengths and the ratios are denoted by $l_{i,q}$ and $r_{k,q}$ respectively, and $q = 1, \ldots, 143$ denotes the index of a member of our cadre family (see Section 3.4). Using the values in Table 4, we compute the matrix $\mathbf{C} = [c_{ij}]$ of the absolute value of the ratio correlation as

$$c_{ij} = \left| \frac{\mu[(r_{i,q} - \mu(r_i))(r_{j,q} - \mu(r_j))]}{\sigma(r_j)\sigma(r_j)} \right|, \quad i, j = 1, \ldots, 28.$$

DEFINITION 3.1. Let $\mathbf{V} = [v_{ki}]$ be a 28×8 matrix. For all $r_k \in \mathcal{R}$, and $l_i \in \mathcal{L}$ we define the following functions:

- $\text{weight}(r_k) = \dfrac{\sigma(r_k)}{\mu(r_k)}$

The variance $\sigma(r_k)$ is an indication of the precision of the statistical information concerning the ratio r_k. Therefore, when the weight of a ratio is small, the ratio is more constrained.

- $\text{cover}(r_k, l_i) = \begin{cases} 1 & \text{if } (r_k = \frac{l_m}{l_n} \wedge (l_i = l_m \vee l_i = l_n)) \\ c_{kb} & \text{otherwise,} \end{cases}$

where $c_{kb} = \max_f \{c_{kf}\}$, and $(r_f = \frac{l_d}{l_e} \wedge (l_i = l_d \vee l_i = l_e)$. The value of $\text{cover}(r_k, l_i)$ measures to what extent the ratio r_k constrains the length l_i.

- $\text{degree}(r_k, \mathbf{V}) = \sum_i v_{ki}$, where $v_{ki} = \text{cover}(r_k, l_i)$.

The degree function measures the correlation of a ratio with all the segments.

- $\text{goodness}(r_k, \mathbf{V}) = \dfrac{\text{degree}(r_k, \mathbf{V})}{\text{weight}(r_k)}$.

The goodness function is employed in determining which ratios will be used to constrain the estimation process as explained in Algorithm 1.

DEFINITION 3.2. A set $\mathcal{B} \subset \mathcal{R}$ (the set of ratios) is a *covering set* of $\mathcal{L}$ (the set of segments), if $\forall l_i \in \mathcal{L}, \exists r_k = l_m / l_n \in \mathcal{R}$ such that $l_i = l_m \vee l_i = l_n$.

Our objective is to find a set of ratios that constrain all segment lengths. Thus, we formulate the problem as a *set covering* problem as follows: If $\mathcal{L}$ is the set of a SM's segment lengths and $\mathcal{R}$ is the set of the corresponding ratios, find the covering set $\mathcal{B}$ for $\mathcal{L}$. In the following, we outline the steps of the algorithm.

ALGORITHM 1 (RATIO SELECTION).

1. $\mathcal{B} := \emptyset$
2. $\forall (r_k, l_i) \in (\mathcal{R} \times \mathcal{L})$, $\mathbf{V}[\mathrm{k, i}] := cover(r_k, l_i)$
3. $\forall i, l_i \in \mathcal{L}$, $\mathrm{care}(l_i) := 0$
4. **while** (True) **do**
5. $\quad \forall j, r_j \in \mathcal{R} \setminus \mathcal{B}$, $d[j] = degree(r_j, \mathbf{V})$
6. $\quad \forall j, r_j \in \mathcal{R} \setminus \mathcal{B}$, $g[j] = goodness(r_j, \mathbf{V})$
7. $\quad m := \arg \max_{j, r_j \in \mathcal{R} \setminus \mathcal{B}} \{g[j]\}$ and $r_m = \frac{l_d}{l_e}$
8. $\quad \mathcal{B} := \mathcal{B} \cup \{r_m\}$, $care[d] += \mathbf{V}[m, d]$, $care[e] += \mathbf{V}[m, e]$
9. $\quad$ **if** $(care[i] \geq 1), \forall_i, l_i \in \mathcal{L}$
10. $\quad$ **then**
11. $\quad\quad$ return $\mathcal{B}$
12. $\quad$ **else**
13. $\quad\quad \forall j, r_j \in \mathcal{R} \setminus \mathcal{B}$, $\mathbf{V}[j, d] := \max\{0, \mathbf{V}[j, d] - care[d]\}$
14. $\quad\quad \forall j, r_j \in \mathcal{R} \setminus \mathcal{B}$, $\mathbf{V}[j, e] := \max\{0, \mathbf{V}[j, e] - care[e]\}$
15. **end while**

The resulting set is $\mathcal{B} = \{\frac{\mathrm{LC}}{\mathrm{UT}+\mathrm{LT}}, \frac{\mathrm{LLA}}{\mathrm{LUA}}, \frac{\mathrm{LHP}}{\mathrm{LUA}}, \frac{\mathrm{LF}}{\mathrm{LUL}}, \frac{\mathrm{LF}}{\mathrm{LLL}}\}$.

PROPOSITION 3.2. *The Ratio Selection algorithm* (*outlined above*) *always returns a covering set.*

Proof. The set $\mathcal{R}$ is the maximum covering set of $\mathcal{L}$. In the worst case, the values of goodness$(r_j, \mathbf{V})$ could all be equal. However, since step 7 in Algorithm 1 returns one index, the corresponding ratio is added to $\mathcal{B}$ and discarded from $\mathcal{R}$, and therefore the algorithm always returns a covering set. ■

4. ANTHROPOMETRY AND POSE ESTIMATION

Our technique for simultaneously estimating the anthropometry and the pose from a single uncalibrated image has the following steps [5]:

ALGORITHM 2 (ANTHROPOMETRY AND POSE ESTIMATION).

Step 1: Selection of projected landmarks
Step 2: Initial anthropometric estimates
Step 3: Initial pose estimates
Step 4: Iterative minimization over lengths and angles

In the following subsections we describe each step in detail.

4.1. Selection of Projected Landmarks

We have developed a simple user interface that allows the user to select the projection of visible landmarks of the subject's body (see Fig. 6a). In addition, the user marks the segments whose orientation is almost parallel to the image plane. For example, in Fig. 6a the green dots depict projection of landmarks associated with segments whose orientation

is almost parallel to the image plane, and the blue dots depict all other selected landmarks. Although information from both types of landmarks will be used for pose estimation, initial length estimates will be based on the projected length of the segments whose orientation is almost parallel to the image plane only.

4.2. Initial Anthropometric Estimates

Our basic assumption is that there is a number of segments whose orientation is almost parallel to the image plane and therefore we can obtain good approximation ratios for them using Proposition 3.1. Thus, our algorithm cannot handle images like the one depicted in Fig. 1b, since one cannot locate segments that are almost parallel to the image plane to obtain reliable initial anthropometric estimates.

Let h_i be the projected length of a segment i on the image, and let $\mathcal{I} \subset \{1, \ldots, 8\}$ be the index set of these segments whose orientation is almost parallel to the image plane. Using the measurements h_i ($i \in \mathcal{I}$), we compute all possible ratios s_k that correspond to the ratios of SMs in Table 4 as follows,

$$s_k = \begin{cases} \frac{h_m}{h_n} & \text{if } \mu(l_n) > \mu(l_m) \\ \frac{h_n}{h_m} & \text{otherwise,} \end{cases}$$

where $k \in \mathcal{K} = \{k \in \{1, \ldots, 28\} | k = n - m(m-1)(8 - \frac{m}{2}), m, n \in \mathcal{I}\}$. Basing our selection on these ratios, we select one SM from the family of 143 SMs whose length ratios closely match the ratios computed from the image using the Mahalanobis distance. To accomplish this goal, we determine

$$q^* = \arg\min_q \sum_{k \in \mathcal{K}} (r_{k,q} - s_k) \left(\sum_{j \in \mathcal{K}} v_{kj} (r_{j,q} - s_j) \right),$$

where $r_{k,q}, r_{j,q}$ are defined in Table 4, $q = 1, \ldots, 143$,

$$v_{kj} = \mu[(r_{k,q} - \mu(r_k))(r_{j,q} - \mu(r_j))]$$

are covariance coefficients of the ratios, and $k, j \in \mathcal{K}$.

$$[v_{kj}] = \mathbf{O}^{-1}$$

and $\mathbf{O}$ is the covariance matrix of the ratios $\{r_{k,q}\}_{k \in \kappa, j = 1, \ldots, 143}$.

The length measurements of the selected q^* stick model are used as initial segment length estimates.

To facilitate the overall understanding of our algorithm, we first present the fourth step in the next section, and then we present the third step in Section 4.4.

4.3. Iterative Minimization over Lengths and Angles

The variables we want to estimate are the lengths of the body segments and their pose. Therefore, we will solve a system of equations where prior information about the human body (e.g., relations between lengths of segments) will provide constraints to an optimization that minimizes the discrepancy between the synthesized appearance of the SM (for that pose) and the image data of the subject in the given image.

As mentioned earlier, the user selects a set of points on the image that correspond to the projection of the sites of the stick model. For each of these points, we set up a point-to-line

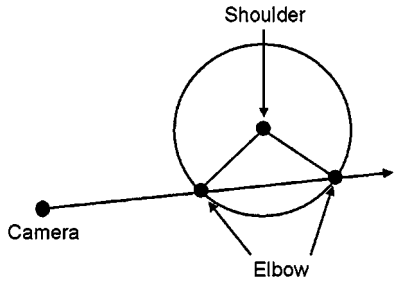

FIG. 5. Two initial poses for the SM's upper arm.

constraint, since the site will lie on a line that goes through the center of the camera and the projection of a landmark. Let $\mathbf{c}$ be the camera's center of projection, $\mathbf{m}_i$ be the position of $m_i \in \mathcal{S}$ (a SM's site), and $\mathbf{m}_i^p$ be the corresponding projection point selected by the user. Then the constraint line is given by $\mathbf{c}_i = \mathbf{c} + \lambda \mathbf{d}_i$, where $\mathbf{d}_i = (\mathbf{m}_i^p - \mathbf{c})/\|\mathbf{m}_i^p - \mathbf{c}\|$. Our objective function is $O = \sum_i \text{distance}(\mathbf{m}_i, \mathbf{c}_i)$. We seek to minimize the value of this function using a BFGS nonlinear solver [22]. Due to the large number of degrees of freedom, in order not to be trapped into a local minimum and to obtain an anthropometric plausible correct answer, we apply the solver in a hierarchical manner. Statistical information about the proportions of the human body and the range of motion of each joint are integrated into the hierarchical optimization method as a set of constraints.

Hierarchical solver. First, to facilitate and expedite the minimization process, we assign a priority to each joint and end effector, and we schedule our optimization to proceed in a hierarchical manner starting with joints closer to the waist joint moving outward. The priorities for each joint are detailed in the column named PR in Table 2.

Constraints. Three classes of constraints are applied: (1) constraints derived from the joint limit information associated with the range of motion of a joint, (2) constraints that enforce the symmetry between the left and right sides of the subject (e.g., the length of the left upper arm is equal to the length of the right upper arm), and (3) constraints that enforce proportions. For the symmetry constraints in particular, we require that the ratios $\{\frac{\text{LY}}{\text{RY}}, \frac{\text{LC}}{\text{RC}}, \frac{\text{LUA}}{\text{RUA}}, \frac{\text{LLA}}{\text{RLA}}, \frac{\text{LHD}}{\text{RHD}}, \frac{\text{LHP}}{\text{RHP}}, \frac{\text{LUL}}{\text{RUL}}, \frac{\text{LLL}}{\text{RLL}}, \frac{\text{LF}}{\text{RF}}\}$ are within ϵ distance from the value one. Thus, the variables whose values will be estimated are the lengths of the segments outlined in Table 3. Furthermore, the constraint for a ratio $r_k = l_m/l_n \forall r_k \in \mathcal{B}(\mathcal{B} = \{\frac{\text{UT+LT}}{\text{LC}}, \frac{\text{LUA}}{\text{LLA}}, \frac{\text{LUA}}{\text{LHP}}, \frac{\text{LUL}}{\text{LP}}, \frac{\text{LLL}}{\text{LP}}\}$ as per Algorithm 1) takes the form

$$l_j - \sigma_j \leq l_j \leq l_j + \sigma_j$$
$$r_k = \frac{l_m}{l_n},$$

TABLE 5
Accuracy of the Length Estimates for the Synthetic Experiment

	$\frac{\text{LC}}{\text{UT+LT}}$	$\frac{\text{LLA}}{\text{LUA}}$	$\frac{\text{LHP}}{\text{LUA}}$	$\frac{\text{LF}}{\text{LUL}}$	$\frac{\text{LF}}{\text{LLL}}$
Actual	0.6553	0.9829	0.5700	0.6397	0.6341
Estimated	0.6517	0.9517	0.5713	0.6595	0.6460
PE %	0.5494	3.1743	0.2281	3.0952	1.8767

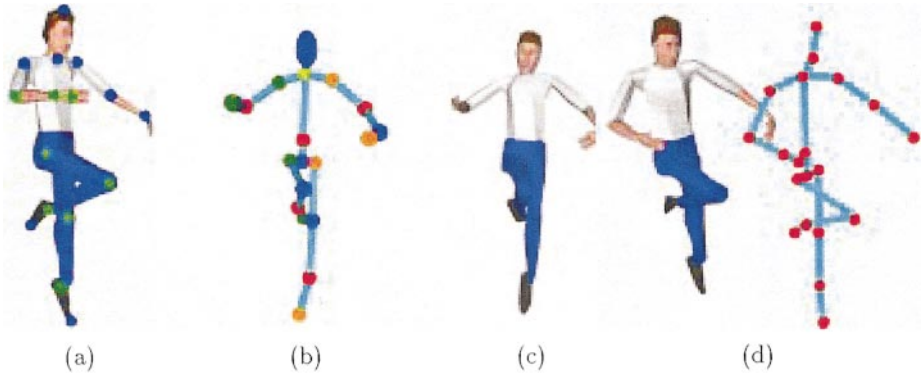

FIG. 6. Synthetic experiment: (a) input image and selected points, (b)–(d) novel views of the reconstructed 3D SM and the virtual human model.

where l_j is either l_m or l_n and corresponds to the segment with the smaller value between μ_m/σ_m and μ_n/σ_n.

4.4. Initial Pose Estimates

In order for the nonlinear solver not to get trapped into a local minimum, we use a geometric method for providing an initial guess for the pose of the segments whose both endpoints were selected by the user. Let $\mathbf{m}_i^p$ be the projection of a site m_i in the image, $l_i > 0$ be the length of the segment of which this landmark is the end-effector, and $\mathbf{j} \in \mathcal{J}$ be the position of the parent joint of that landmark on the stick model. By construction, the following equation applies,

$$\|\mathbf{c} + \lambda \mathbf{d}_i - \mathbf{j}\| = l_i,$$

where $\mathbf{d}_i$ is the unit direction between the camera and $\mathbf{m}_i^p$, as defined earlier. This quadratic equation has two solutions

$$\lambda_1 = \mathbf{d}_i \cdot (\mathbf{j} - \mathbf{c}) + \sqrt{[\mathbf{d}_i \cdot (\mathbf{c} - \mathbf{j})]^2 - \|\mathbf{c} - \mathbf{j}\|^2 + l_i^2} \quad \text{and}$$

$$\lambda_2 = \mathbf{d}_i \cdot (\mathbf{j} - \mathbf{c}) - \sqrt{[\mathbf{d}_i \cdot (\mathbf{c} - \mathbf{j})]^2 - \|\mathbf{c} - \mathbf{j}\|^2 + l_i^2},$$

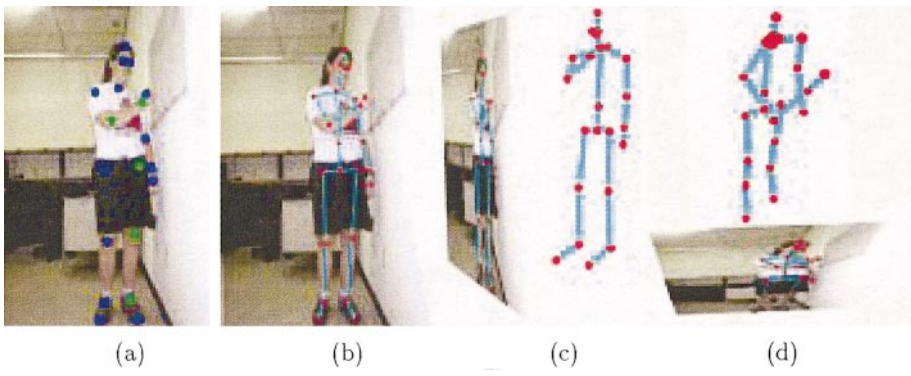

FIG. 7. Subject Vanessa: (a) selected points, (b) reconstructed model overlaid to the image, and (c), (d) novel views of the reconstructed SM.

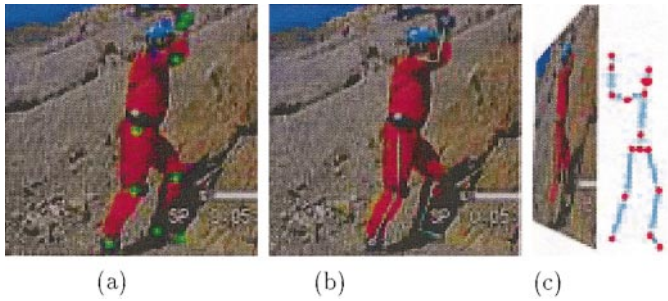

FIG. 8. (a) Input image depicting a geologist along with the user-selected input landmarks, (b) reconstructed model overlaid to the image, and (c) novel view of the reconstructed SM.

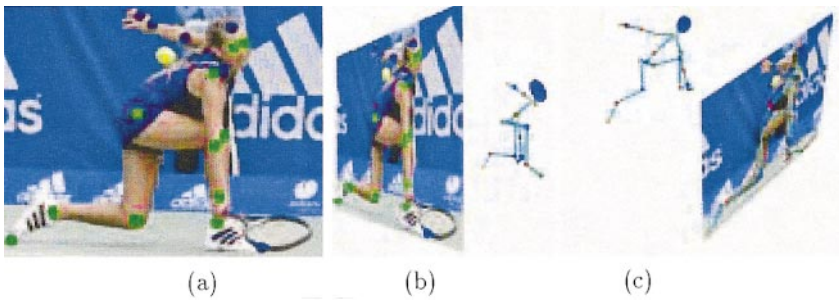

FIG. 9. (a) Input image depicting a tennis player along with the user-selected input landmarks, (b) reconstructed model overlaid to the image, and (c) novel view of the reconstructed SM.

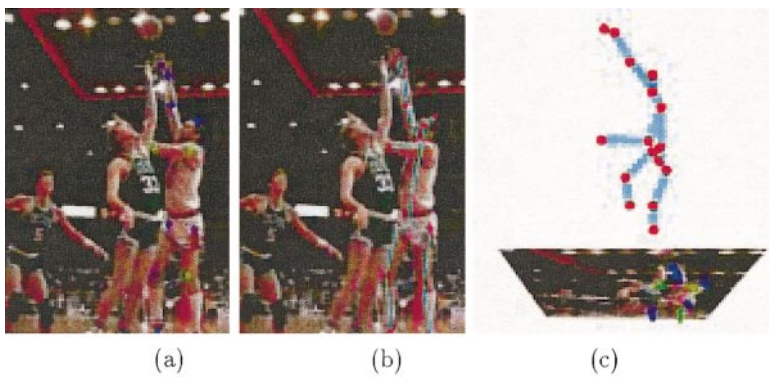

FIG. 10. (a) Input image depicting a basketball player along with the user-selected input landmarks, (b) reconstructed model overlaid to the image, and (c) novel view of the reconstructed SM.

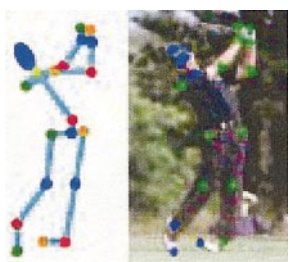

FIG. 11. Input image depicting a golf player along with the user-selected input landmarks, and the reconstructed model SM.

TABLE 6
Accuracy of the Pose Estimates for the Synthetic Experiment

Joint	Actual	Estimated	PE %
at	(0.00°, 0.00°, 0.00°)	(0.00°, 0.00°, 0.00°)	0.00
sp	(−0.50°, −2.00°, 0.00°)	(−0.48°, −1.98°, 0.00°)	1.37
la	(12.97°, −15.37°, −63.40°)	(13.60°, −16.50°, −65.00°)	3.09
lc	(12.24°, 0.99°)	(12.01°, 0.99°)	1.87
le	(41.84°)	(42.31°)	1.12
lh	(−16.80°, 2.14°, −2.73°)	(−16.53°, 2.12°, −2.69°)	1.60
lk	(0.00°)	(0.0°)	0.00
ls	(2.50°, 36.47°, 3.41°)	(2.50°, 36.87°, 3.37°)	1.09
lw	(42.67°, −39.12°, 1.85°)	(43.02°, −40.12°, 1.65°)	1.86
ra	(32.28°, 33.54°, −50.13°)	(31.98°, 33.32°, −49.81°)	0.72
rc	(11.70°, 11.41°)	(11.61°, 11.51°)	0.82
re	(62.74°)	(63.05°)	0.49
rh	(−8.63°, −13.40°, 63.87°)	(−8.62°, −13.40°, 63.77°)	0.15
rk	(129.13°)	(131.02°)	1.46
rs	(26.20°, 31.11°, 42.44°)	(26.20°, 31.07°, 43.01°)	0.97
rw	(−5.35°, −6.55°, −54.53°)	(−5.15°, −7.05°, −54.50°)	0.98
wt	(0.00°, 0.00°, 0.00°)	(0.00°, 0.00°, 0.00°)	0.00

that correspond to the intersection of the line $\mathbf{c} + \lambda_k \mathbf{d}_i$ with the sphere of radius l_i centered at $\mathbf{j}$. For example, Fig. 5 depicts the two possible solutions related to the joint *rs* and the site *re*. The two possible initial guesses for the position of site m_i are $\mathbf{m}_{i1} = \mathbf{c} + \lambda_1 \mathbf{d}_i$ and $\mathbf{m}_{i2} = \mathbf{c} + \lambda_2 \mathbf{d}_i$. Finally, joint limit information is used to prune the solutions that are not feasible. If both positions are feasible, then they are used as initial values for the nonlinear solver.

5. RESULTS AND DISCUSSION

We have performed a number of experiments on synthetic and real data to assess the accuracy, limitations, and advantages of our approach. In all the examples of the input images, the green dots depict projection of landmarks associated with segments whose orientation is almost parallel to the image plane, and the blue dots depict all other selected landmarks (see for example the input image depicted in Fig. 7a). In the first experiment, we applied our technique to an image created using the virtual human modeling tool EAI Jack. Figure 6a depicts the selected points in the input image, while Fig. 6b depicts the reconstructed 3D model. Figures 6c and 6d depict the reconstructed 3D model in novel views. Tables 5 and 6 contain statistical information related to the accuracy of the estimation process.

TABLE 7
Accuracy of the Length Estimates for the Subject *Vanessa*

	$\frac{\text{LC}}{\text{UT}+\text{LT}}$	$\frac{\text{LLA}}{\text{LUA}}$	$\frac{\text{LHP}}{\text{LUA}}$	$\frac{\text{LF}}{\text{LUL}}$	$\frac{\text{LF}}{\text{LLL}}$
Actual	0.6279	0.8625	0.6949	0.5517	0.4778
Estimated	0.6402	0.8516	0.6728	0.5594	0.4888
PE %	1.9589	1.2638	3.1803	1.3957	2.3022

In the second experiment, we applied our technique to a real image from the subject Vanessa whose anthropometric dimensions were manually measured. Figure 7a depicts the selected points, Fig. 7b depicts the reconstructed model overlaid to the image, and Figs. 7c and 7d depict the reconstructed model from novel views. Table 7 captures the percentage error (PE) in estimating the length ratios. We observe that the estimation of anthropometric information is within 3.2% of the anthropometric dimensions of the subject. In general, we have performed numerous other experiments with a variety of subjects whose anthropometric dimensions are known with similar very encouraging results.

In the third experiment, we applied our algorithm to a variety of images from a variety of application domains, where anthropometric information about the subjects was not available. Figures 8a, 9a, and 10a depict the input images along with the selected points, while Figs. 8b and 8c, 9b and 9c, and 10b and 10c depict the reconstructed models from different viewpoints. Finally, Fig. 11 depicts the input image of golf player along with the reconstructed three-dimensional stick model.

6. CONCLUSION

In this paper, we have described a four-step technique for generating anthropometric and posture information for a human subject from a single image. The user initially selects a set of image points that constitute the projection of selected landmarks. Based on the image coordinates of the selected points and anthropometric statistics, pose and anthropometric measurements are obtained by minimizing an appropriate cost function subject to the associated constraints. The novelty of our approach is the use of anthropometric statistics to constrain the estimation process that allows the simultaneous estimation of both anthropometry and pose. We have demonstrated the accuracy, advantages, and limitations of our method for various classes of both synthetic and real input data.

ACKNOWLEDGMENTS

We thank Dr. D. Sivakumar for many useful discussions and Honda R&D Americas, Inc. for their financial support for this work.

REFERENCES

1. A. Azarbayejani, C. Wren, and A. Pentland, Real-time 3-D tracking of the human body, in *Proc. of the IMAGE'COM 96, Bordeaux, France, May 1996.*
2. F. Azuola, *Error in the Representation of Anthropometric Data By Human Figure Models*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1996.
3. N. I. Badler, C. B. Phillips, and B. L. Webber, *Simulating Humans: Computer Graphics Animation and Control*, Oxford Univ. Press, New York, NY, 1993.
4. C. Bregler and J. Malik, Tracking people with twists and exponential maps, in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Santa Barbara, 1998*, pp. 8–15.
5. C. Barrón and I. Kakadiaris, Estimating anthropometry and pose from a single image, in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hilton Head, SC, 2000*, pp. I:669–676.
6. T. J. Cham and J. Rehg, A multiple hypothesis approach to figure tracking, in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, 1999*, pp. 239–245.

7. Q. Delamarre and O. Faugeras, 3D articulated models and multi-view tracking with silhouettes, in *Proc. of the 7th International Conference on Computer Vision, Kerkyra, Greece, September 20–27, 1999*, pp. 716–721.
8. I. Douros, L. Dekker, and B. F. Buxton, An improved algorithm for reconstruction of the surface of the human body from the 3D scanner data using local B-spline patches, in *IEEE International Workshop on Modeling People, Corfu, Greece, September 20, 1999*, pp. 29–36.
9. D. M. Gavrila and L. S. Davis, 3-D model-based tracking of humans in action: A multiview approach, in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, 1996*, pp. 73–80.
10. L. Goncalves, E. Bernardom, E. Ursella, and P. Perona, Monocular tracking of the human arm in 3D, in *Proc. of the Fifth International Conference on Computer Vision, Boston, June 20–22, 1995*, pp. 764–770.
11. J. Gu, T. Chang, I. Mak, S. Gopalsamy, H. C. Shen, and M. F. Yuen, A 3D reconstruction system for human body modeling, in *Proc. First Int. Workshop on Modeling and Motion Capture Techniques for Virtual Environments (CAPTECH 98), Geneva Switzerland, November 26–27, 1998* (N. Magnenat-Thalmann and Daniel Thalmann, Eds.), pp. 229–241.
12. A. Hilton, Towards model-based capture of a person's shape, appearance and motion, in *IEEE International Workshop on Modeling People, Corfu, Greece, September 20, 1999*, pp. 37–44.
13. S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, S. Kawato, K. Ebihara, and S. Morishima, Real-time, 3D estimation of human body postures from trinocular images, in *IEEE International Workshop on Modeling People, Corfu, Greece, September 20, 1999*, pp. 3–10.
14. I. A. Kakadiaris and D. Metaxas, Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection, in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, 1996*, pp. 81–87.
15. I. A. Kakadiaris and D. Metaxas, 3D Human body model acquisition from multiple views, *Int. J. Comput. Vision* **30**, 1998, 191–218.
16. I. A. Kakadiaris and D. Metaxas, Model-based estimation of 3D human motion, in *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12), 2000, 1453–1459.
17. H. J. Lee and Z. Chen, Determination of 3D human body postures from a single view, *Comput. Vision Graphics Image Process.* **30**, 1985, 148–168.
18. National Aeronautics and Space Administration, Man systems integration standards, Technical report, 1987.
19. R. Plankers, P. Fua, and N. D'Apuzzo, Automated body modeling from video sequences, in *IEEE International Workshop on Modeling People, Corfu, Greece, September 20, 1999*, pp. 45–52.
20. J. M. Rehg and T. Kanade, Model-based tracking of self-occluding articulated objects, in *Proc. of the Fifth International Conference on Computer Vision, Boston, 1995*, pp. 612–617.
21. S. Wachter and H.-H. Nagel, Tracking of persons in monocular image sequences, in *Proc. of IEEE Nonrigid and Articulated Motion Workshop, Puerto Rico, June 16, 1997*, pp. 2–9.
22. J. Zhao and N. I. Badler, Nonlinear programming for highly articulated figures, *ACM Trans. Graphics* **13**, 1994, 313–336.