第 29 卷 第 3 期
2003 年 5 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 29, No. 3
May, 2003

Human Body Tracking Using EM Based on a 2-D Articulated Body Model[1)]

Yu Huang Thomas S. Huang
(*Beckman Institute, UIUC, Urbana, Illinois, IL*61801, *U. S. A.*)
(E-mail: {yuhuang, huang}@ifp. uiuc. edu)

Abstract Visual tracking of human body movement is a key technology in a number of areas, such as visual surveillance and monitoring. In this paper we present a 2-D model-based method of human body tracking from a monocular video sequence. Morris & Rehg put forward a 2-D scaled prismatic model(SPM) for figure registration which has far fewer singularity problems than 3-D models. Here we extend it in a 2-D cardboard human body model with additional one DOF of width change. Based on this modified 2-D model rather than 3-D model in Bregler & Malik's work, we also set up a mixture motion model for body movements and then solve motion parameters of the articulated body using EM in a statistical framework, where the model-based kinematic constraints are incorporated in a linear form. Tracking results from real video sequences are encouraging.

Key words Body tracking, expectation-maximization, 2-D articulated motion

1 Introduction

Because of many potentially important applications, "Looking at People" is currently one of the most active application domains in computer vision[1,2]. This trend is motivated by a wide spectrum of applications, such as smart visual surveillance and monitoring, virtual reality, HCI, content-based video indexing, model-based image coding and video conferencing. From a technical point of view, this domain is rich and challenging because of the need to segment rapidly changing scenes in natural environments involving non-rigid motion and camera projection singularity, large number of DOFs, (self) occlusion and image noises from background and human clothing.

The human body models can be simple as the articulated body skeleton with connected segments[3,4] or 2-D patches-based models[5~7], and even be complex as volumetric models with combinations of elementary volumes[8~10]. The statistical techniques have been exploited to infer a generic template form from a labeled training set associated with representative deformation models[11]. Some works are put on learning the human body movement using PCA[8,12], and the learned body motion modes, such as walking, running or jumping, will help model matching during on-line body tracking.

Methods for full body tracking typically use sparse cues such as background difference images, color and edges[13]. Motion or optic flow give rich information, but can cause the tracking model to "drift off" the target[5,6]. The use of template avoids this problem, but template tracking is sensitive to changes in view and illumination[14]. Multiple camera views are often employed to reduce ambiguity and problems due to self-occlusion[7,15,16]. Recently some researchers have regarded body tracking as an inference problem[17,18], they adopt the Bayesian formulation and estimate the model parameters over time using sampling-based particle filtering and multiple hypothesis tracking[3,8,9].

However, it is pointed out in [2]: A 2-D approach is effective for applications where

1) Supported partially by the NSF(CDA96-24396, EIA-99-75019 and IIS-00-85980)
Received October 08, 2002; in revised form April 10, 2003
收稿日期 2002-10-08; 收修改稿日期 2003-04-10

precise pose recovery is not needed or possible due to low image resolution, also with a single human involving constrained movement and single viewpoint. A 3-D approach makes more sense for application in indoor environments where one desires a high-level of discrimination between various unconstrained and complex (multiple) human movements, leading to a more accurate, compact representation of physical space which allows a better prediction and handling of occlusion and collision. The benefit of using multiple cameras to achieve tighter 3-D pose recovery has been quite evident, only the added calibration effort is worthwhile. It also has reduced kinematic singularities if omitting some particular configurations of the articulated object. Unfortunately, in some certain tracking applications, we have only single video source available such as movie footage.

1.1 Related work

Ju defined a cardboard person model, where a person's limbs are represented by a set of connected patches[5]. The image motion of these patches was constrained to enforce articulated motion and solved for from direct estimation of parameterized flow models. Unfortunately, these kinematic constraints cannot be incorporated into the motion estimation framework in a linear form, and meanwhile they did not yet consider the occlusion problem.

Morris mainly analyzed the singularity problem occurred in 3-D articulated object tracking[14]. Morris in fact considered the direct 3-D tracking task decomposed as two separate goals: one is the registration objective, another is the reconstruction goal[14]. Therefore 2-D model-based tracking just was to solve the alignment of 3-D model projection with the image features. They proposed a 2-D scaled prismatic model(SPM) for figure registration, which has far fewer singularity problems than 3-D models. The proposed SPM acts in a plane parallel to the plane of the camera and simulate the image motion of the 3-D model. Each SPM link can rotate around an axis perpendicular to the image plane, and meanwhile it translates along this axis. Eventually, they employed an SSD-based registration method for body tracking.

Bregler used twists and exponential maps to model the 3-D articulated motion[6]. It results in solving simple linear systems to recover robustly the kinematic DOFs under scaled orthogonal projection. They used the idea of support maps to deal with self-occlusion and solve it by the EM algorithm. They had extended it to multiple camera views.

Howe realized human body tracking from single camera video[12]. They also used a 2-D cardboard body model and also performed a task similar to [14] based on the tracking algorithm in [5], only the occlusion problem was taken into account through the use of support maps similar to [6]. Furthermore, they reconstructed the 3-D motion based on the 2-D tracking result and prior model of 3-D model learned from training data.

1.2 Review

In this paper, we propose a human body tracking method from a monocular video. Our body model is like a 2-D cardboard model[5], only its kinematic constraints are generated in a similar way as [14]. Morris assumes that a template is attached to each link which rotates and scales with the link[14]. The rotational DOF captures the effect on the link orientation, and the translational DOF models the foreshortening that occurs when 3D links rotate into and out of the image plane. But we can observe that when the link rotates into and out of the image plane the template width also has changed. So in this paper, we consider further a translational component in the direction orthogonal to the link axis, which will effectuate extension of the planar patch in width. Based on a mixture motion model for body movement, we solve motion parameters of the articulated body in a statistical framework using the EM algorithm, where the kinematic constraints can be incorporated in a linear form. E-

ventually, encouraging tracking results of human frontal and lateral walking are also given.

2 Framework of articulated body tracking

A general framework for model-based tracking consists of four main components involved: prediction, synthesis, image analysis and state estimation. Different approaches have been discussed in this framework. One possibility is to use a "divide-and-conquer technique" where an articulated object is decomposed into a number of primitive (rigid or articulated) sub-parts; one solves for motion and depth of the sub-parts and verifies whether the parts satisfy the necessary constraints. Instead other approaches use parameterized models where the articulation constraints are encoded in the representation itself, thus they take advantage as much as possible of prior knowledge and rely as little as possible on error prone 2-D image segmentation. This group can be divided into two types of methods: one is using such a parameterized model to update poses by inverse kinematics; another does not attempt to inverse a non-linear measurement equation, instead it uses the measurement equation directly to synthesize the model and then uses a fitting measure between synthesized and observed features for feedback. Our method belongs to the former one in this group.

2.1 Dominant motion estimation

Assuming the changes in image intensity are only due to translation of local image intensity, the inter-frame motion is defined as

$$f(\boldsymbol{x}, t+1) = f(\boldsymbol{x} - \boldsymbol{u}(\boldsymbol{x};\boldsymbol{a}), t) \tag{1}$$

with $f(\boldsymbol{x}, t)$ as the brightness function in time instant t, $\boldsymbol{x}=(x, y)$ as the coordinate of the image pixel, and $\boldsymbol{u}(\boldsymbol{x};\boldsymbol{a})$ as the motion vector. Without loss of generality, we select affine transform as the motion model

$$\boldsymbol{u}(\boldsymbol{x};\boldsymbol{a}) = \begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} = \begin{bmatrix} a_0 + a_1\, x + a_2\, y \\ a_3 + a_4\, x + a_5\, y \end{bmatrix}$$

where $\boldsymbol{a}=(a_0, a_1, a_2, a_3, a_4, a_5)^{\mathrm{T}}$ is the affine motion vector.

The M-estimator is employed to solve the affine motion of body patch R,

$$\min_{(u,v)} E_D = \sum_{(x,y)\in R} \rho(u f_x + v f_y + f_t, \sigma) \tag{2}$$

where f_x, f_y, f_t are partial derivatives of brightness function with respect to x, y and t, and the ρ- function can be chosen as the Geman-McClure function

$$\rho(x,\sigma) = \frac{x^2}{x^2 + \sigma^2}$$

with σ as the scale parameter.

To solve the problem, the iterative weighted least squares (IWLS) method is performed to find robustly the motion parameters[19], which solves a nonlinear LS problem iteratively,

$$\begin{bmatrix} \sum wf_x^2 & \sum wxf_x^2 & \sum wyf_x^2 & \sum wf_x f_y & \sum wxf_x f_y & \sum wyf_x f_y \\ \sum wxf_x^2 & \sum wx^2 f_x^2 & \sum wxyf_x^2 & \sum wxf_x f_y & \sum wx^2 f_x f_y & \sum wxyf_x f_y \\ \sum wyf_x^2 & \sum wxyf_x^2 & \sum wy^2 f_x^2 & \sum wyf_x f_y & \sum wxyf_x f_y & \sum wy^2 f_x f_y \\ \sum wf_x f_y & \sum wxf_x f_y & \sum wyf_x f_y & \sum wf_y^2 & \sum wxf_y^2 & \sum wyf_y^2 \\ \sum wxf_x f_y & \sum wx^2 f_x f_y & \sum wxyf_x f_y & \sum wxf_y^2 & \sum wx^2 f_y^2 & \sum wxyf_y^2 \\ \sum wyf_x f_y & \sum wxyf_x f_y & \sum wy^2 f_x f_y & \sum wyf_y^2 & \sum wxyf_y^2 & \sum wy^2 f_y^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} = \begin{bmatrix} -\sum wf_x f_t \\ -\sum wxf_x f_t \\ -\sum wyf_x f_t \\ -\sum wf_t f_y \\ -\sum wxf_t f_y \\ -\sum wyf_t f_y \end{bmatrix}$$

where $w(r)=\psi(r)/r$, with derivative $\psi(r)=\mathrm{d}\rho(r)/\mathrm{d}r$ and error $r=uf_x+vf_y+f_t$. The estimate of σ is given by a robust measure as

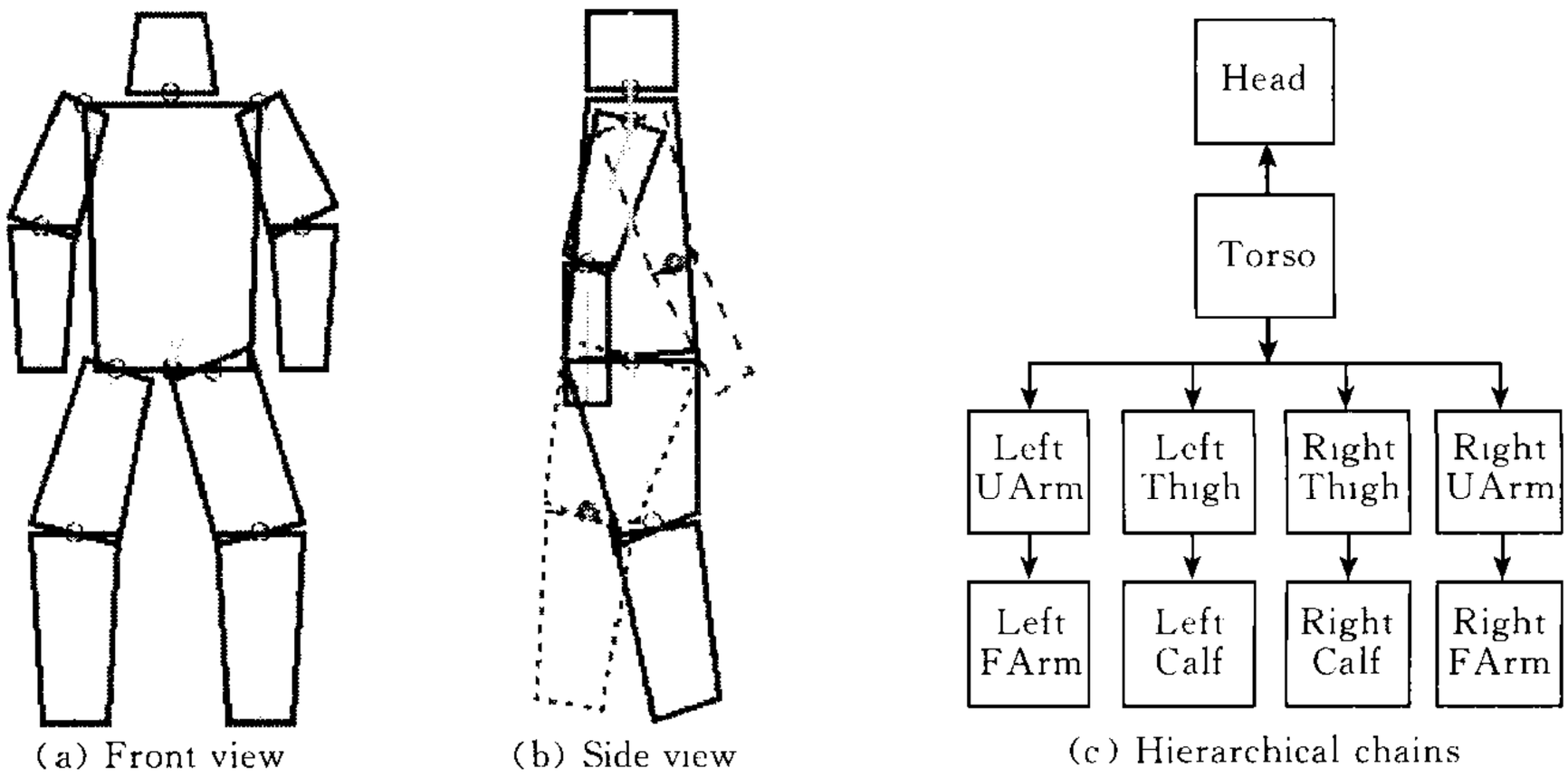
$$\sigma = 1.4826 median_i |r_i|$$

The algorithm begins with constructing the Gaussian pyramid (we set up three lev-

els). At the coarse level motion is initially set to zero. The number of iterations is chosen as 10. When the estimated parameters are interpolated into the next level, these parameters are used to warp (realized by bilinear interpolation) the last frame to the current frame. In the current level only the change in the paramenters are estimated in the iterative update scheme.

2.2 Human body model

In Fig. 1 the defined 2-D articulated cardboard body model is illustrated. Its kinematics is similar to that of the 2-D PSM[14]. We define each body part as an isosceles trapezoidal planar patch either from the front view Fig. 1(a) or from the side view Fig. 1(b), and those patches are linked together by the kinematic chains in a hierarchical manner shown in Fig. 1(c). In the front view and side view, the joints are labeled in yellow and the links are illustrated in red, the origin of the body coordinate system is located at the joint of the torso depicted in black in Fig. 1(a), overlapped with the joint of thigh in Fig. 1(b). Here because of occlusion, we only consider half of the body model in the side view (the dashed lines depict the occluded body parts).



(a) Front view (b) Side view (c) Hierarchical chains

Fig. 1 Human body model

In the tracking initialization, users need to determine which view suitable for the scenario, and meanwhile the joint locations and angles, width and length of every planar patch are determined easily. Below we will discuss in detail the kinematic constraints tc the body model.

2.3 Articulated motion estimation

An articulated object's posture can be parameterized in many different ways. One possibility is to use a set of parameters such as the position and orientation parameters of each part and impose the constraints in a Lagrangian form[5]. Another approach is to use the kinematic chain equations and select parameters, such as the orientations of each part and the position of a reference point[6,16]. Here we go the latter way.

Feature attributes are expressed as a function of the kinematic state variable, i. e. joint angles. Here for an image pixel $I_j(\boldsymbol{q})$ to the 3-D point $\boldsymbol{p}_j$ we have the forward kinematic equation given the state vector $\boldsymbol{q}$,

$$I_j(\boldsymbol{q}) = I(P\boldsymbol{F}(\boldsymbol{q}, \boldsymbol{p}_j))$$

with 3-D kinematics by the nonlinear function $\boldsymbol{F}(\boldsymbol{q}, \boldsymbol{p}_j)$ under orthographic projection P. The image velocity for pixel $I_j(\boldsymbol{q})$ can be expressed as

$$V_p = J \cdot \dot{\boldsymbol{q}}, \quad J_{ji} = (\nabla I_j)^{\mathrm{T}} P \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{q}}(\boldsymbol{q}_0, \boldsymbol{p}_j)$$

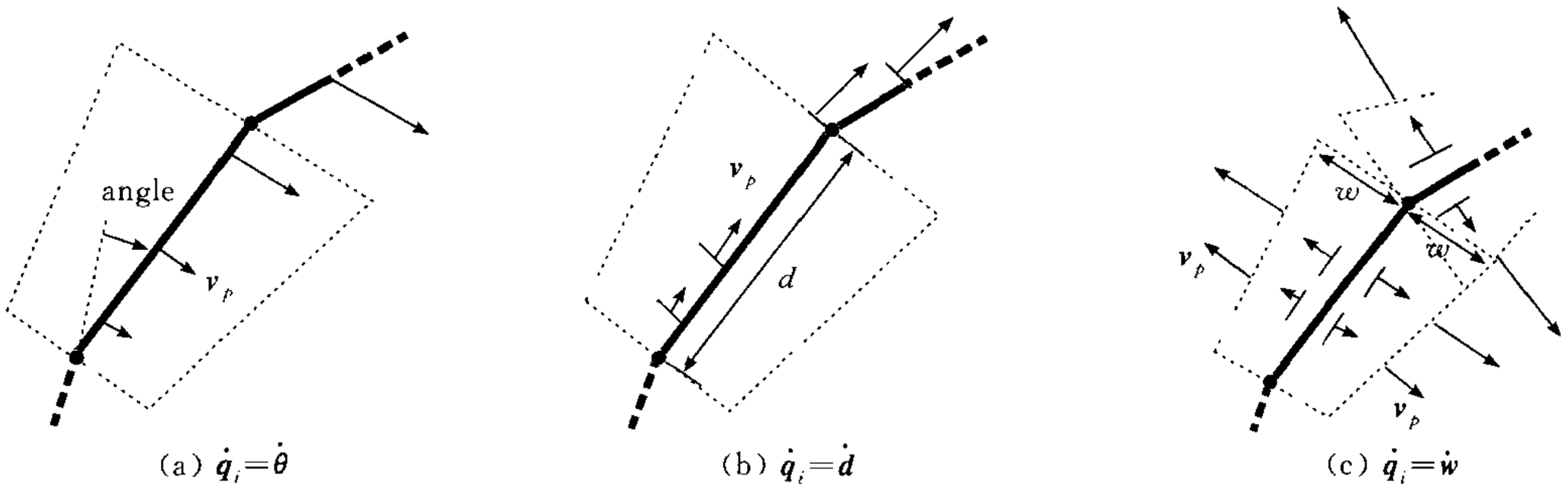
with 3-D kinematic Jacobian $J_j^k = \partial \boldsymbol{F}/\partial \boldsymbol{q}$ and the image gradient ∇I_j. By defination[20], we have $\dot{\boldsymbol{p}}_j = J_j^k \dot{\boldsymbol{q}}$.

Since a column of the Jacobian J_i maps the state velocity $\dot{q}_i$ to an image velocity, by finding the image velocity in terms of this state we can obtain an expression for J_i. Here we make the same assumptions as [14]: The Jacobian is effected by linear combination of each state independently. Consequently, we derive the Jacobian form as below.

1) If $\dot{\boldsymbol{q}}_i = \dot{\boldsymbol{\theta}}$ is the angle of a revolute joint, it will contribute an angular velocity component to links further along the chain given by $\dot{w} = \dot{q}\boldsymbol{a}$, where $\boldsymbol{a}$ is the axis of rotation (the z axis). The image velocity, $\boldsymbol{v}_p$ shown in Fig. 2(a), of a point at location $\boldsymbol{r}$ on the chain resulting from this rotation is given by: $\boldsymbol{v}_p = P\boldsymbol{\omega} \times \boldsymbol{r} = P\boldsymbol{a} \times \boldsymbol{r}\dot{q} = \boldsymbol{r}_{2d}\dot{q}$. So, we get the Jacobian J_i

$$J_{ik}^j = \begin{cases} 0, & \textit{links } k, \text{ where } k < i \\ \boldsymbol{r}_{2d}, & \textit{links } k, \text{ where } k \geqslant i \end{cases}, \quad \boldsymbol{r}_{2d} = \begin{bmatrix} -(y_{jk} - y_{0i}) \\ x_{jk} - x_{0i} \end{bmatrix} \tag{3}$$

with (x_{0i}, y_{0i}) as joint of the link.



(a) $\dot{\boldsymbol{q}}_i = \dot{\boldsymbol{\theta}}$ (b) $\dot{\boldsymbol{q}}_i = \dot{\boldsymbol{d}}$ (c) $\dot{\boldsymbol{q}}_i = \dot{\boldsymbol{w}}$

Fig. 2 The 2-D Scaled Prismatic Planar Model showing image velocities due to state velocities

2) If $\dot{\boldsymbol{q}}_i = \dot{\boldsymbol{d}}$ refers to the extension of the scaled prismatic link, its derivative will contribute a velocity component to points on the link proportional to their position on the link: bq_i, where b is the fractional position of the point over the total length extension q_i. The velocity component for a point on the links is thus $\boldsymbol{v}_p = bq_i\dot{q}_i\boldsymbol{n}_i$ shown in Fig. 2(b). Subsequent links will be effected only by the end-point extension of the link, and so have a velocity component given by $\boldsymbol{v}_p = q_i\dot{q}_i\boldsymbol{n}_i$. Here $\boldsymbol{n}_i$ is the link axis vector. Hence the Jacobian is given by

$$J_{ik}^j = \begin{cases} 0, & \textit{links } k, \text{ where } k < i \\ bq_i\boldsymbol{n}_i, & \textit{link } i \ (0 \leqslant b \leqslant 1) \\ q_i\boldsymbol{n}_i, & \textit{links } k, \text{ where } k > i \end{cases}, \quad q_i\boldsymbol{n}_i = \begin{bmatrix} x_{0i+1} - x_{0i} \\ y_{0i+1} - y_{0i} \end{bmatrix} \tag{4}$$

$$b = \frac{(x_{jk} - x_{0i})(x_{0i+1} - x_{0i}) + (y_{jk} - y_{0i})(y_{0i+1} - y_{0i})}{(x_{0i+1} - x_{0i})^2 + (y_{0i+1} - y_{0i})^2}$$

with (x_{0i+1}, y_{0i+1}) as end of the link axis.

3) If $\dot{\boldsymbol{q}}_i = \dot{\boldsymbol{w}}$ corresponds to width expansion of the model shown in Fig. 2(c), its derivative will generate a velocity component for those points belonging to this link given by $\boldsymbol{v}_p = abq_i\dot{q}_i\boldsymbol{m}_i^{\pm}$, where b is the same as above, a is the fractional position of the point over the total width extension. Here q_i referers to the bottom width of the trapezoidal patch as actually there is no extension for the top width, and $\boldsymbol{m}_i^{\pm}$ represent two unitary vectors orthogonal to the link axis with the contrary directions. Subsequent patches will also be effected, to guarantee the trapezoidal symmetry we approximate them as $\boldsymbol{v}_p = a'q_i\dot{q}_i\boldsymbol{m}_k^{\pm}$, $k > i$, where $\boldsymbol{m}_k^{\pm}$ are a pair of unitary vectors like $\boldsymbol{m}_i^{\pm}$ and a' is calculated the same way as a, corresponding to point (x_{jk}, y_{jk}) attached to link k. Consequently, we get the Jacobian as

$$J_{ik}^{j}=\begin{cases}0, & links\ k,\ k<i\\ abq_i\boldsymbol{m}_i^{-}, & link\ i,\ negative\ direction(0\leqslant a,b\leqslant 1)\\ abq_i\boldsymbol{m}_i^{+}, & link\ i,\ positive\ direction(0\leqslant a,b\leqslant 1)\\ a'q_i\boldsymbol{m}_k^{-}, & link\ k,\ negative\ direction(0\leqslant a'\leqslant 1),\ k>i\\ a'q_i\boldsymbol{m}_k^{+}, & link\ k,\ positive\ direction(0\leqslant a'\leqslant 1),\ k>i\end{cases},\quad q_i\boldsymbol{m}_i^{\pm}=\begin{bmatrix}x_{1^{\pm}i}-x_{0i+1}\\ y_{1^{-}i}-y_{0i+1}\end{bmatrix} \tag{5}$$

$$\text{if } w_{\text{bottom}}>w_{\text{top}},\ a=\frac{(x_{ji}-x_{0i+1})(x_{1^{\pm}i}-x_{0i+1})+(y_{ji}-y_{0i-1})(y_{1^{\pm}i}-y_{0i+1})}{((1-b)w_{\text{top}}+bw_{\text{bottom}})\sqrt{(x_{1^{+}i}-x_{0i+1})^2+(y_{1^{\pm}i}-y_{0i-1})^2}}$$

$$\text{if } w_{\text{bottom}}\approx w_{\text{top}},\ a=\frac{(x_{ji}-x_{0i+1})(x_{1^{+}i}-x_{0i+1})+(y_{ji}-y_{0i+1})(y_{1^{\pm}i}-y_{0i+1})}{(x_{1^{\pm}i}-x_{0i+1})^2+(y_{1^{-}i}-y_{0i+1})^2}$$

with $(x_{1^{-}i},y_{1^{-}i})$ and $(x_{1^{+}i},y_{1^{+}i})$ as a pair of patch corners close to the end of link axis. For concise description, we can define

$$a'q_i\boldsymbol{m}_k^{\pm}=(a'q_i/q_k)q_k\boldsymbol{m}_k^{\pm}=a^{(i)}\begin{bmatrix}x_{1^{+}k}-x_{0k+1}\\ y_{1^{+}k}-y_{0k+1}\end{bmatrix}$$

with (here we can not guarantee $0<a^{(i)}<1$)

$$a^{(i)}=\frac{(x_{jk}-x_{0k+1})(x_{1^{\pm}k}-x_{0k+1})+(y_{jk}-y_{0k+1})(y_{1^{-}k}-y_{0k+1})}{((1-b')w'_{\text{top}}+b'w'_{\text{bottom}})[(x_{1^{\pm}k}-x_{0k+1})^2+(y_{1^{-}k}-y_{0k+1})^2]}\cdot$$
$$\cdot\sqrt{(x_{1^{\pm}i}-x_{0i+1})^2+(y_{1^{+}i}-y_{0i+1})^2}$$

where b' is calculated similarly to b, corresponding to point (x_{jk},y_{jk}) attached to link k. If we make further approximation as $q_i\approx q_k$ (especially when there are only two links in the chain), then we have $a^{(i)}\approx a'$.

Eventually, from three formulas (3~5) we get all the entries of the Jacobian for each pixel I_j in the link i. If we have a chain of $K+1$ segments linked with K joints, then we have

$$\begin{bmatrix}u(x,y)\\ v(x,y)\end{bmatrix}=J\cdot[\dot\theta_1\quad \dot d_1\quad \dot w_1\quad \cdots\quad \dot\theta_K\quad \dot d_K\quad \dot w_K]^{\mathrm{T}}=J\dot{\boldsymbol{q}} \tag{6}$$

Substituting (6) into (2) leads to another IWLS problem as (detailed derivation are given in Appendix)

$$\min_{\dot{\boldsymbol{q}}}E_D=\rho(H\dot{\boldsymbol{q}}+z,\sigma) \tag{7}$$

with $\boldsymbol{z}=[f_{t_{11}}\ \cdots\ f_{t_{12}}\ \cdots\ f_{t_{1K}}\ \cdots]^{\mathrm{T}}$, $H=[H_1\quad H_2\quad \cdots\quad H_K]$, and

$$H_1=\begin{bmatrix}-f_{x_{11}}(y_{11}-y_{01})+ & b_{11}[f_{x_{11}}(x_{02}-x_{01})+ & a_{11}b_{11}[f_{x_{11}}(x_{1^{\pm}1}-x_{02})+\\ f_{y_{11}}(x_{11}-x_{01}) & f_{y_{11}}(y_{02}-y_{01})] & f_{y_{11}}(y_{1^{\pm}1}-y_{02})]\\ -f_{x_{12}}(y_{12}-y_{01})+ & f_{x_{12}}(x_{02}-x_{01})+ & a_{12}^{(1)}[f_{x_{12}}(x_{1^{\pm}2}-x_{03})+\\ f_{y_{12}}(x_{12}-x_{01}) & f_{y_{12}}(y_{02}-y_{01}) & f_{y_{12}}(y_{1^{\mp}2}-y_{03})]\\ \cdots & \cdots & \cdots\\ -f_{x_{1K}}(y_{1K}-y_{01})+ & f_{x_{1K}}(x_{02}-x_{01})+ & a_{1K}^{(1)}[f_{x_{1K}}(x_{1^{\pm}K}-x_{0K+1})+\\ f_{y_{1K}}(x_{1K}-x_{01}) & f_{y_{1K}}(y_{02}-y_{01}) & f_{y_{1K}}(y_{1^{+}K}-y_{0K+1})]\\ \cdots & \cdots & \cdots\end{bmatrix}$$

$$H_2=\begin{bmatrix}0 & 0 & 0\\ \cdots & \cdots & \cdots\\ -f_{x_{12}}(y_{12}-y_{02})+ & b_{12}[f_{x_{12}}(x_{03}-x_{02})+ & a_{12}b_{12}[f_{x_{12}}(x_{1^{\pm}2}-x_{03})+\\ f_{y_{12}}(x_{12}-x_{02}) & f_{y_{12}}(y_{03}-y_{02})] & f_{y_{12}}(y_{1^{\pm}2}-y_{03})]\\ \cdots & \cdots & \cdots\\ -f_{x_{1K}}(y_{1K}-y_{02})+ & f_{x_{1K}}(x_{03}-x_{02})+ & a_{1K}^{(2)}[f_{x_{1K}}(x_{1^{\pm}K}-x_{0K+1})+\\ f_{y_{1K}}(x_{1K}-x_{02}) & f_{y_{1K}}(y_{03}-y_{02}) & f_{y_{1K}}(y_{1^{\pm}K}-y_{0K+1})]\\ \cdots & \cdots & \cdots\end{bmatrix}$$

$$H_K = \begin{bmatrix} 0 & 0 & 0 \\ \cdots & \cdots & \cdots \\ 0 & 0 & 0 \\ \cdots & \cdots & \cdots \\ -f_{x_{1K}}(y_{1K}-y_{0K})+ & b_{1K}[f_{x_{1K}}(x_{0K+1}-x_{0K})+ & a_{1K}b_{1K}[f_{x_{1K}}(x_{1^{\pm}K}-x_{0K+1})+ \\ f_{y_{1K}}(x_{1K}-x_{0K}) & f_{y_{1K}}(y_{0K+1}-y_{0K})] & f_{y_{1K}}(y_{1^{\pm}K}-y_{0K+1})] \\ \cdots & \cdots & \cdots \end{bmatrix}$$

To make the motion estimation more robust, we perform a variable order model fitting strategy. It implies that: we first calculate the parameters of a model with only rotation, then one with rotation and extension along the axis, finally the one with rotation, extension along the axis and in width.

In our experiments, we decompose the human body into different groups, where each group therefore consists of one or two rigid segments:

Front view	Side view
1. torso;	1. torso;
2. head;	2. head;
3～4. left (right) upper arm→left (right) front arm;	3. upper arm→front arm;
5～6. left (right) thigh→left (right) calf.	4. thigh→calf.

Torso motion is special, as the base link we estimate independently its 5-D affine motion: two translational parameters, rotation angles around its joint, and two scaling parameters, i. e.

$$\begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} = \begin{bmatrix} t_x \\ t_y \end{bmatrix} + \begin{bmatrix} s_x\cos\Delta\theta & s_y\sin\Delta\theta \\ -s_x\sin\Delta\theta & s_y\cos\Delta\theta \end{bmatrix} \begin{bmatrix} x-x_o \\ y-y_o \end{bmatrix}$$

If we are given some prior knowledge, for example, the torso only take translational motion or are static, we are able to simplify motion estimation of the torso. As the base link, the torso will effectuate other links, i. e. neck, arms and legs. Thus we apply the estimated torso motion parameters to register the frame before estimating other segments' motion.

2.4 Motion tracking based on the support maps in a statistical framework

It is convenient to assume that a family of motion models takes the form of a Gaussian mixture density. This means that each image measurement is drawn independently from a Gaussian distribution whose mean is a function of motion model parameters. Here our measurements are the image pixels $\boldsymbol{I}=\{I_1, I_2, \cdots, I_N\}$, so we want to assign each pixel to some body part or the background. We let $r_m(\boldsymbol{I}, \dot{\boldsymbol{q}})$ denote the measurement deviation at pixel $\boldsymbol{I}$ with respect to model m (background or body part, here the background is assumed to be static) whose motion is described by $\dot{\boldsymbol{q}}$.

Actually our method is motion segmentation-based, thus possibly some pixels from the environment will distract our motion measurements. To alleviate this difficulty, one available way is to resort to background modeling and subtraction so that the labeling region is constraint into the foreground pixels. Some modern methods introduce the spatial proximity or coherence[21～23] like "blobs" grouping in [13, 24]. Therefore, the total likelihood for a pixel measurement $\boldsymbol{I}$ at (x, y) can be written:

$$L(\boldsymbol{I}, x, y \mid \dot{\boldsymbol{q}}) = \sum_{m=0}^{K} \frac{\pi_m}{\sigma_m} e^{-r_m^2(\boldsymbol{I}, \dot{\boldsymbol{q}})/2\sigma_m^2} P_m(x, y \mid \dot{\boldsymbol{q}})$$

with the spatial proximity prior $P_m(x, y \mid \dot{\boldsymbol{q}})$ as Gaussian distribution (position means and covariance) for body parts, i. e.

$$P_m(x,y \mid \dot{\boldsymbol{q}}) = P_m(\boldsymbol{x} \mid \dot{\boldsymbol{q}}) = \frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_m)^{\mathrm{T}} K_m^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_m)}}{2\pi\sqrt{|K_m|}},\ \boldsymbol{\mu}_m = E[\boldsymbol{x}],\ K_m = E[(\boldsymbol{x}-\boldsymbol{\mu}_m)(\boldsymbol{x}-\boldsymbol{\mu}_m)^{\mathrm{T}}]$$

and uniform distribution for the background. The variance σ_m^2 controls the softness of the partition, while the mixture probabilities π_m describe the likelihood of assigning $\boldsymbol{I}$ to model m such that

$$\sum_{m=1}^{K} \pi_m = 1 \tag{8}$$

At a local extrema[25], it can be shown that π_m and $\dot{\boldsymbol{q}}$ must satisfy

$$\pi_m = \frac{\sum_{i=1}^{N} \tau_{mi}}{N},\ m = 1,2,\cdots,K \tag{9}$$

$$\min_{\dot{\boldsymbol{q}}} E_D = \sum_m \sum_i \tau_{mi}\rho(H\dot{\boldsymbol{q}} + \boldsymbol{z}, \sigma) \tag{10}$$

Here the quantities τ_{mi} represent the ownership probabilities, or support maps, calculated by

$$\tau_{mi} = \frac{\frac{\pi_m}{\sigma_m} e^{-r_m^2(I_i)/2\sigma_m^2} P_m(x,y \mid \dot{\boldsymbol{q}})}{\sum_{t=0}^{K} \frac{\pi_t}{\sigma_t} e^{-r_t^2(I_i)/2\sigma_t^2} P_m(x,y \mid \dot{\boldsymbol{q}})},\ m = 1,2,\cdots,K;\ i = 1,2,\cdots,N \tag{11}$$

where the solution σ_m^2 is given by

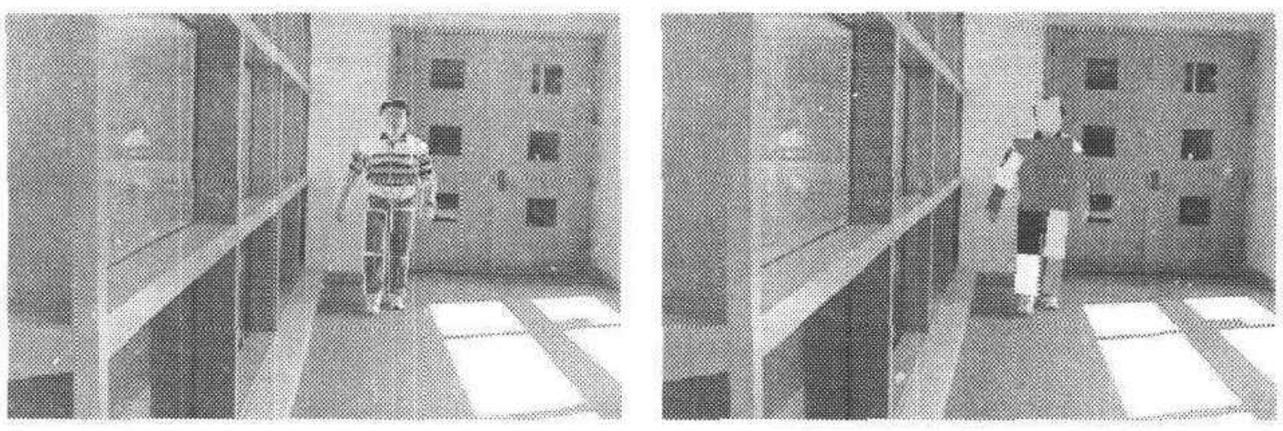
$$\hat{\sigma}_m^2 = \frac{\sum_{i=1}^{N} \tau_{mi} r_m^2(I_i)}{N\pi_m},\ m = 1,2,\cdots,K \tag{12}$$

We can find that, the Gaussian spatial priors (for human body patches) die off quickly beyond a certain region of interest for pruning. Meanwhile, we update them by re-calculating means and covariance for support layers. Eventually, the support maps are used to weight our IWLS problem in Section 2.3 to calculate $\dot{\boldsymbol{q}}$.

The EM (Expectation-Maximization) algorithm is based on a general statistical framework for estimating mixture model parameters from incomplete, or missing data. In the case of motion analysis the missing data is the segmentation which assigns each image measurement to a motion model. Now we construct the EM framework for object tracking based on the support-layered representation of environments. We start with an initial guess of the support maps. 1) (M-step) calculate the motion parameters $\dot{\boldsymbol{q}}$ using formula (10); 2) (E-step) given $\dot{\boldsymbol{q}}$, adjust the state of the body model, determine the spatial prior and recalculate the support maps at the next frame.

3 Experiment results

Several sequences are captured in the lab corridor for testing. We realize the whole algorithm in VC++ 6.0 at PC. Currently the processing speed on Pentium II 400M is about 25s/frame. When initializing the 2-D body model in the first image frame, we assume the occlusion order is known: for example, in Fig. 3(a) the left(right) arm will occlude the torso and the left(right) thigh, and in Fig. 3(b) the left arm possibly occlude the torso and the left thigh while the right side is not considered during tracking. We move our 2-D model's origin, joints and four corners for each planar patch to fit the body view in the image, so the initial support map for each body part can be determined. Next the state variables $\boldsymbol{q}$ for the body posture are calculated, i.e. joint angles, lengths and widths of body parts (For convenience, we first fix origin of the torso, then the joint of each part, and finally the four corners).

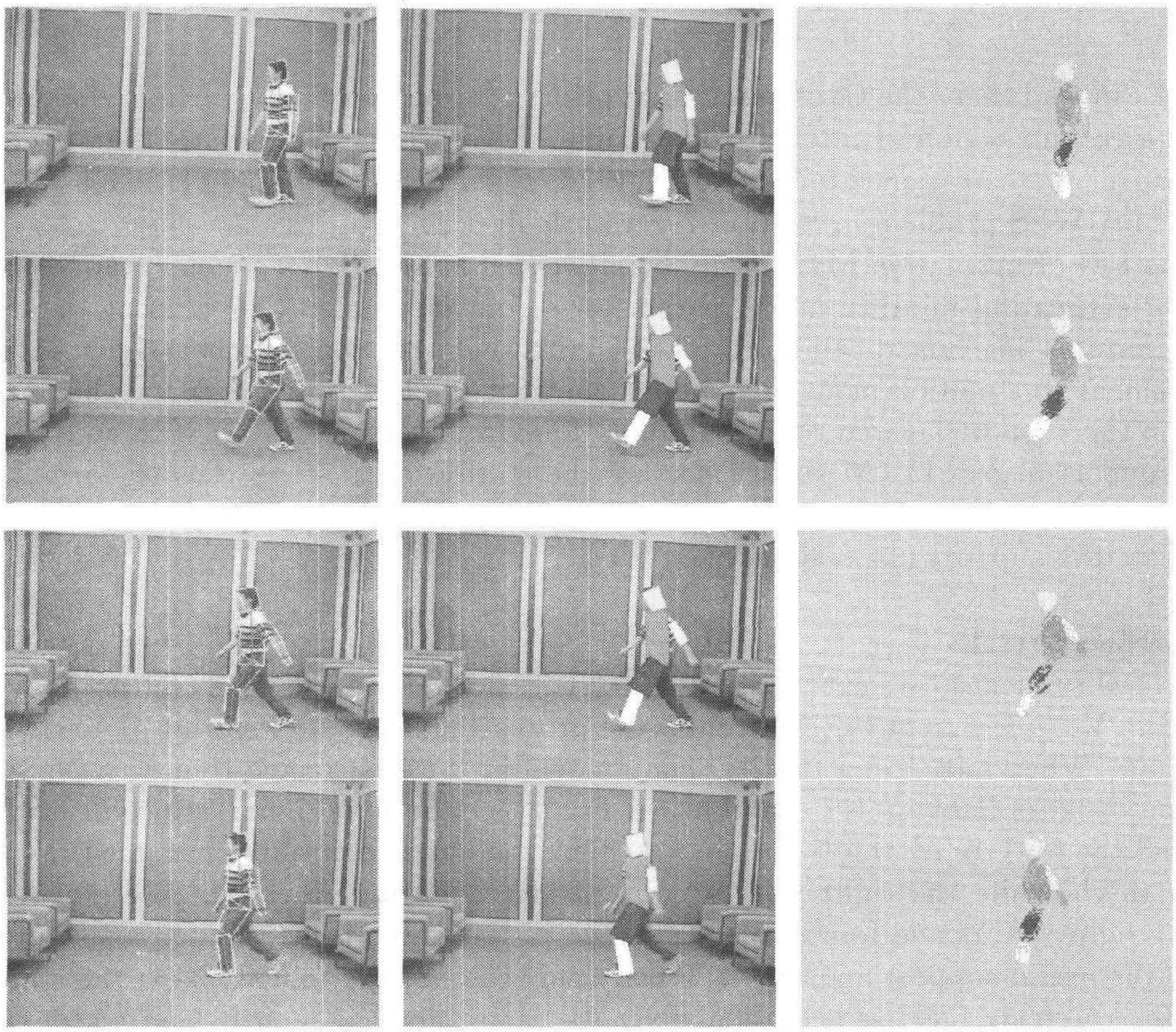


(a) Front model fitting

(b) Side model fitting

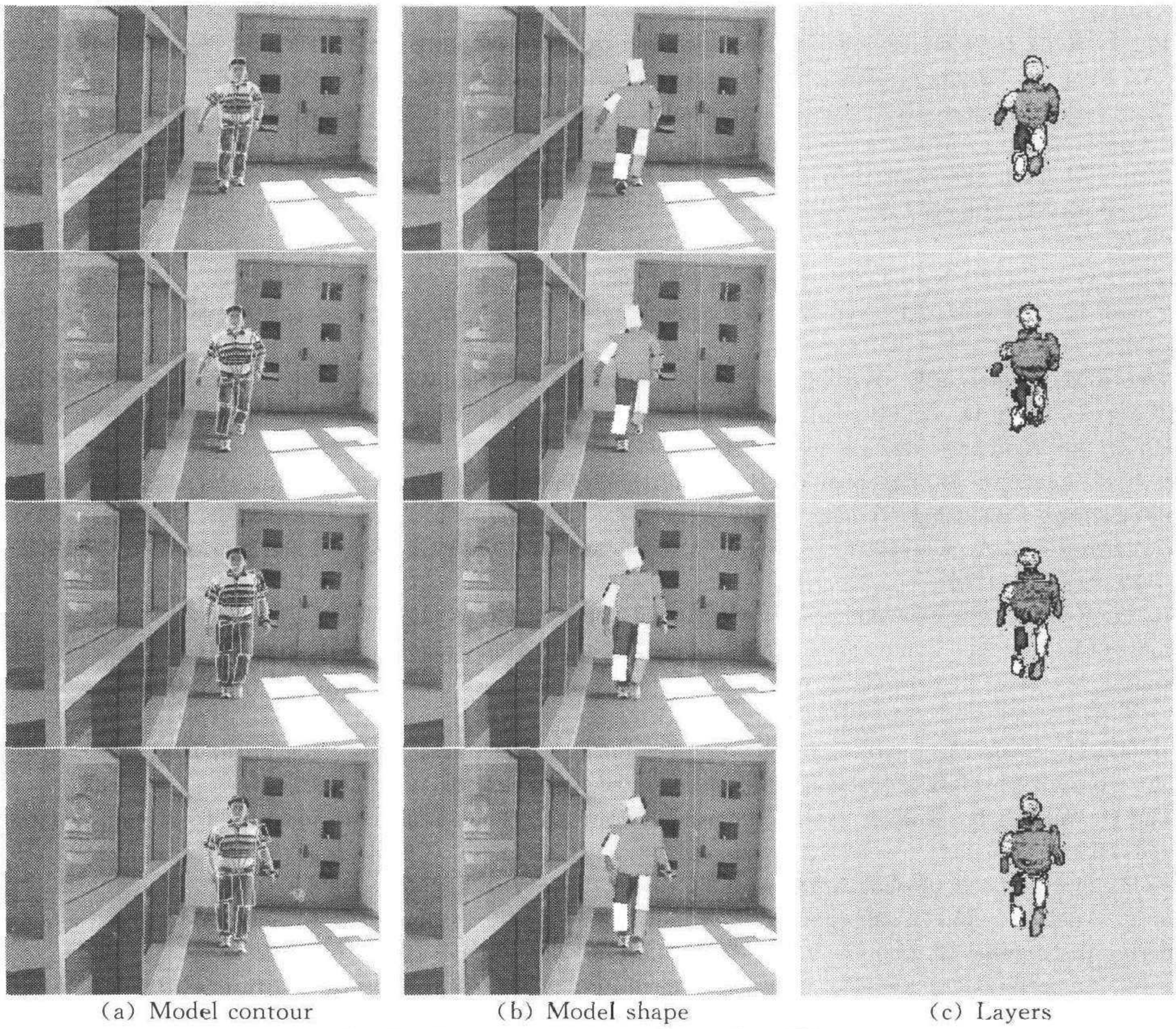
Fig. 3 Initialization of model-based body tracking

Two tracking results for lateral and frontal walking respectively are given in Fig. 4 and Fig. 5 respectively, each body part of the model based on the estimated state or postures are illustrated in different colors. We haven't ground-truth for 3-D body motion and camera geometry, the performance could be evaluated from the images overlapped by the 2-D model. In Fig. 4, body movements of lateral walking almost happen to be in the image



(a) Model contour (b) Model shape (c) Layers

Fig. 4 Body tracking of lateral walking



(a) Model contour (b) Model shape (c) Layers

Fig. 5 Body tracking of frontal walking

plane, so we don't observe obvious foreshortening. But for frontal walking in Fig. 5, we could notice the distinct foreshortening both in length and width. These tracking results are encouraging.

4 Conclusion and future work

We realize a human body tracking method from a monocular video sequence based on a 2-D cardboard model. We discuss its kimematics which can capture the 3-D motion of the human body. Besides of the rotation around the joint and translation along the axis in [14], we add a translational component in the direction orthogonal to the link axis, which will effectuate extension of the planar patch in width. Compared with Ju's work[5], we set up kinematic constraints in a linear form. Based on a mixture motion model for body movement, we solve motion parameters of the articulated body in a statistical framework using the EM algorithm, where the kinematic constraints are also incorporated. Experiment results of our method from both lateral and frontal walking sequences are encouraging.

Based on a learned 3-D prior model from motion captured data by PCA we could construct a Bayesian framework to infer 3-D reconstruction from 2-D tracking results[12]. We also want to learn some constraints on the single arm or leg movement only, so a hierarchical prior body motion model could be set up for use.

References

1 Aggarwal J, Cai Q. Human body analysis: A review. In: IEEE Workshop on Non-rigid and Articulated Motion, 1997. 90～102

2 Gavrila D M. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*,

1999, **73**(1)
3 Cham T-J, Rehg J. A multiple hypothesis approach to figure tracking. In: CVPR'99, 1999. 239～245
4 Song Y, Feng X, Perona P. Towards detection of human motion. In: CVPR'00, 2000
5 Ju S X, Black M J, Yacoob Y. Cardboard people: A parameterized model of articulated image motion. In: International Conference on Automatic Face and Gesture Recognition, 1996. 38～44
6 Bregler C, Malik J. Tracking people with twists and exponential maps. In: CVPR'98, 1998. 8～15
7 Jojic N, Turk M, Huang T S. Tracking self-occluding articulated objects in density disparity maps. In: ICCV'99, 1999. 123～130
8 Duetscher J, Blake A, Reid I. Articulated motion capture by annealed particle filtering. In: CVPR'00, 2000
9 Sidenbladh H, Black M, Fleet D. Stochastic tracking of 3D human figures using 2D image motion. In: ECCV'00, 2000. 702～718
10 Yamamoto M, Yagishita K. Scene constraints-aided tracking of human body. In: CVPR'00, 2000. 151～256
11 Baumberg A, Hogg D. An efficient method for contour tracking using active shape models. In: IEEE Workshop on Motion of Non-rigid and Articulated Objects, 1994. 194～199
12 Howe N R, Leventon M, Freeman W. Bayesian reconstruction of 3D human motion from single-camera video. *Neural Information Processing Systems*, 1999
13 Wren Azarbayejani A, Darrell T, Pentland A. Pfinder: Real-time tracking of the human body. *IEEE T-PAMI*, 1997, **19**(7): 780～785
14 Morris D, Rehg J. Singular analysis for articulated object tracking. In: IEEE, CVPR'98, 1998
15 Delamarre Q, Faugeras O. 3D articulated models and multi-view tracking with silhouettes. In: ICCV'99, 1999
16 Yamamoto M, Sato A, Kawada S. Incremental tracking of human actions from multiple views. In: CVPR'98, 1998. 2～7
17 Sidenbladh H, Black M. Learning image statistics for Bayesian tracking. In: ICCV'01, 2001
18 Ioffe S, Forsyth D. Human tracking with mixtures of trees. In: ICCV'01, 2001
19 Holland P, Welsch R. Robust regression using iteratively reweighted least squares. In: Comm. Statist. Theor. Methods, 1977
20 Asada H, Slotine J J E. Robot Analysis and Control. New York: Wiley, 1986
21 Weiss Y, Adelson E H. A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models. In: CVPR'96, 1996. 321～326
22 Weiss Y. Smoothness in Layers: Motion segmentation using nonparametric mixture estimation. In: CVPR'97, 1997. 520～527
23 Vasconcelos N, Lippman A. Empirical Bayesian EM-based motion segmentation. In: CVPR'97
24 Bregler C. Learning and recognizing human dynamics in video sequences. In: CVPR'97
25 Mclachlan G J, Basford K E. Mixture Models Inference and Applications to Clustering. New York: Markel Dekker Inc., 1988

Appendix

We define symmetric matrix $G_{kk} = H_k^{\mathrm{T}} H_k$ with these components (g_{st}^k, $s=1,2,3, t=1,2,3$), $k=1,\cdots,K$, i. e.

$$g_{11}^k = \sum_{i\geqslant k}\sum_j [-f_{x_{ji}}(y_{ji}-y_{0k})+f_{y_{ji}}(x_{ji}-x_{0k})]^2$$

$$g_{12}^k = g_{21}^k = \sum_{i>k}\sum_j [-f_{x_{ji}}(y_{ji}-y_{0k})+f_{y_{ji}}(x_{ji}-x_{0k})][f_{x_{ji}}(x_{0(k+1)}-x_{0k})+f_{y_{ji}}(y_{0(k+1)}-y_{0k})] + \sum_j b_{jk}[-f_{x_{jk}}(y_{jk}-y_{0k})+f_{y_{jk}}(x_{jk}-x_{0k})][f_{x_{jk}}(x_{0(k+1)}-x_{0k})+f_{y_{jk}}(y_{0(k+1)}-y_{0k})]$$

$$g_{22}^k = \sum_{i>k}\sum_j [f_{x_{ji}}(x_{0(k+1)}-x_{0k})+f_{y_{ji}}(y_{0(k+1)}-y_{0k})]^2 + \sum_j b_{jk}^2[f_{x_{jk}}(x_{0(k+1)}-x_{0k})+f_{y_{jk}}(y_{0(k+1)}-y_{0k})]^2$$

$$g_{13}^k = \sum_j a_{jk}b_{jk}[-f_{x_{jk}}(y_{jk}-y_{0k})+f_{y_{jk}}(x_{jk}-x_{0k})][f_{x_{jk}}(x_{1\pm k}-x_{0k+1})+f_{y_{jk}}(y_{1\pm k}-y_{0k+1})] + \sum_{i>k}\sum_j a_{ji}^{(k)}[-f_{x_{ji}}(y_{ji}-y_{0k})+f_{y_{ji}}(x_{ji}-x_{0k})][f_{x_{ji}}(x_{1\pm i}-x_{0i+1})+f_{y_{ji}}(y_{1\pm i}-y_{0i+1})]$$

$$g_{23}^k = g_{32}^k = \sum_j a_{jk}b_{jk}^2[f_{x_{jk}}(x_{1\pm k}-x_{0k+1})+f_{y_{jk}}(y_{1\pm k}-y_{0k+1})][f_{x_{jk}}(x_{0(k+1)}-x_{0k})+f_{y_{jk}}(y_{0(k+1)}-y_{0k})] + \sum_{i>k}\sum_j a_{ji}^{(k)}[f_{x_{ji}}(x_{1\pm i}-x_{0i+1})+f_{y_{ji}}(y_{1\pm i}-y_{0i+1})][f_{x_{ji}}(x_{0(k+1)}-x_{0k})+f_{y_{jk}}(y_{0(k+1)}-y_{0k})]$$

$$g_{33}^k = \sum_j a_{jk}^2 b_{jk}^2[f_{x_{jk}}(x_{1\pm k}-x_{0k+1})+f_{y_{jk}}(y_{1\pm k}-y_{0k+1})]^2 + \sum_{i>k}\sum_j (a_{ji}^{(k)})^2[f_{x_{ji}}(x_{1\pm i}-x_{0i+1})+f_{y_{ji}}(y_{1\pm i}-y_{0i+1})]^2$$

Next, we define matrix $G_{lm} = H_l^{\mathrm{T}} H_m$ with each component as g_{st}^{lm}, $s=1,2,3, t=1,2,3$, $m\neq l$, $l=1,\cdots,K$,

$m=1,\cdots,K$. Without loss of generality we set $l>m$, thus

$$g_{11}^{lm}=\sum_{i\geqslant l}\sum_{j}[-f_{x_{ji}}(y_{ji}-y_{0l})+f_{y_{ji}}(x_{ji}-x_{0l})][-f_{x_{ji}}(y_{ji}-y_{0m})+f_{y_{ji}}(x_{ji}-x_{0m})]$$

$$g_{12}^{lm}=\sum_{i\geqslant l}\sum_{j}[-f_{x_{ji}}(y_{ji}-y_{0l})+f_{y_{ji}}(x_{ji}-x_{0l})][f_{x_{ji}}(x_{0(m+1)}-x_{0m})+f_{y_{ji}}(y_{0(m+1)}-y_{0m})]$$

$$g_{13}^{lm}=\sum_{i\geqslant l}\sum_{j}a_{ji}^{(m)}[-f_{x_{ji}}(y_{ji}-y_{0l})+f_{y_{ji}}(x_{ji}-x_{0l})][f_{x_{ji}}(x_{1^{\pm}i}-x_{0i+1})+f_{y_{ji}}(y_{1^{\pm}i}-y_{0i+1})]$$

$$g_{21}^{lm}=\sum_{i>l}\sum_{j}[f_{x_{ji}}(x_{0(l+1)}-x_{0l})+f_{y_{ji}}(y_{0(l+1)}-y_{0l})][-f_{x_{ji}}(y_{ji}-y_{0m})+f_{y_{ji}}(x_{ji}-x_{0m})]+$$
$$\sum_{j}b_{jl}[f_{x_{jl}}(x_{0(l+1)}-x_{0l})+f_{y_{jl}}(y_{0(l+1)}-y_{0l})][-f_{x_{jl}}(y_{jl}-y_{0m})+f_{y_{jl}}(x_{jl}-x_{0m})]$$

$$g_{22}^{lm}=\sum_{i>l}\sum_{j}[f_{x_{ji}}(x_{0(l+1)}-x_{0l})+f_{y_{ji}}(y_{0(l+1)}-y_{0l})][f_{x_{ji}}(x_{0(m+1)}-x_{0m})+f_{y_{ji}}(y_{0(m+1)}-y_{0m})]+$$
$$\sum_{j}b_{jl}[f_{x_{jl}}(x_{0(l+1)}-x_{0l})+f_{y_{jl}}(y_{0(l+1)}-y_{0l})][f_{x_{jl}}(x_{0(m+1)}-x_{0m})+f_{y_{jl}}(y_{0(m+1)}-y_{0m})]$$

$$g_{23}^{lm}=\sum_{j}a_{jl}^{(m)}b_{jl}[f_{x_{jl}}(x_{0(l+1)}-x_{0l})+f_{y_{jl}}(y_{0(l+1)}-y_{0l})][f_{x_{jl}}(x_{1^{\pm}l}-x_{0l+1})+f_{y_{jl}}(y_{1^{\pm}l}-y_{0l+1})]+$$
$$\sum_{i>l}\sum_{j}a_{ji}^{(m)}[f_{x_{jl}}(x_{0(l+1)}-x_{0l})+f_{y_{jl}}(y_{0(l+1)}-y_{0l})][f_{x_{ji}}(x_{1^{\pm}i}-x_{0i+1})+f_{y_{jl}}(y_{1^{\pm}i}-y_{0i+1})]$$

$$g_{31}^{lm}=\sum_{j}a_{jl}b_{jl}[f_{x_{jl}}(x_{1^{\pm}l}-x_{0l+1})+f_{y_{jl}}(y_{1^{\pm}l}-y_{0l+1})][-f_{x_{ji}}(y_{ji}-y_{0m})+f_{y_{ji}}(x_{ji}-x_{0m})]+$$
$$\sum_{i>l}\sum_{j}a_{ji}^{(l)}[f_{x_{ji}}(x_{1^{\pm}i}-x_{0i+1})+f_{y_{ji}}(y_{1^{\pm}i}-y_{0i+1})][-f_{x_{ji}}(y_{ji}-y_{0m})+f_{y_{ji}}(x_{ji}-x_{0m})]$$

$$g_{32}^{lm}=\sum_{j}a_{jl}b_{jl}[f_{x_{jl}}(x_{1^{\pm}l}-x_{0l+1})+f_{y_{jl}}(y_{1^{+}l}-y_{0l+1})][f_{x_{jl}}(x_{0(m+1)}-x_{0m})+f_{y_{jl}}(y_{0(m+1)}-y_{0m})]+$$
$$\sum_{i>l}\sum_{j}a_{ji}^{(l)}[f_{x_{ji}}(x_{1^{\pm}i}-x_{0i+1})+f_{y_{ji}}(y_{1^{\pm}i}-y_{0i+1})][f_{x_{ji}}(x_{0(m+1)}-x_{0m})+f_{y_{ji}}(y_{0(m+1)}-y_{0m})]$$

$$g_{33}^{lm}=\sum a_{jl}^{(m)}a_{jl}b_{jl}[f_{x_{jl}}(x_{1^{\pm}l}-x_{0l+1})+f_{y_{jl}}(y_{1^{\pm}l}-y_{0l+1})]^2+$$
$$\sum_{i>l}\sum_{j}a_{ji}^{(l)}a_{ji}^{(m)}[f_{x_{ji}}(x_{1^{\pm}i}-x_{0i+1})+f_{y_{ji}}(y_{1^{\pm}i}-y_{0i+1})]^2$$

So we have $G=H^{\mathrm{T}}H$ with each component G_{st}, $s=1,\cdots,K$, $t=1,\cdots,K$. Consequently, we define $\boldsymbol{c}=-H^{\mathrm{T}}\boldsymbol{z}$. For convenience we represent as $C=[\boldsymbol{c}_1\quad\boldsymbol{c}_2\quad\cdots\quad\boldsymbol{c}_K]^{\mathrm{T}}$, where $\boldsymbol{c}_k=-H_k^{\mathrm{T}}\boldsymbol{z}$ $(k=1,\cdots,K)$ as

$$\boldsymbol{c}_k=\begin{bmatrix}\sum_{i\geqslant k}\sum_{j}-f_{t_{ji}}[-f_{x_{ji}}(y_{ji}-y_{0k})+f_{y_{ji}}(x_{ji}-x_{0k})]\\ \sum_{i>k}\sum_{j}-f_{t_{ji}}[f_{x_{ji}}(x_{0(k+1)}-x_{0k})+f_{y_{ji}}(y_{0(k+1)}-y_{0k})]-\\ \sum_{j}b_{jk}f_{t_{jk}}[f_{x_{jk}}(x_{0(k+1)}-x_{0k})+f_{y_{jk}}(y_{0(k+1)}-y_{0k})]\\ \sum_{i>k}\sum_{j}-a_{ji}^{(k)}f_{t_{ji}}[f_{x_{ji}}(x_{1^{\pm}i}-x_{0i+1})+f_{y_{ji}}(y_{1^{\pm}i}-y_{0i+1})]+\\ \sum_{j}-a_{jk}b_{jk}f_{t_{jk}}[f_{x_{jk}}(x_{1^{\pm}k}-x_{0k+1})+f_{y_{jk}}(y_{1^{\pm}k}-y_{0k+1})]\end{bmatrix}$$

Therefore we can solve the IWLS problem

$$G_w\dot{\boldsymbol{q}}=\boldsymbol{c}_w.$$

where G_w and $\boldsymbol{c}_w$ are weighted version of G and $\boldsymbol{c}$ respectively by their support maps.

Yu Huang Received his bachelor degree in Xi'an Jiaotong University of P. R. China, Department of Information & Control Engineering, in 1990; master degree in Xidian University of P. R. China, Department of Electrical Engineering, in 1993; Ph. D. degree in Northern Jiaotong University of P. R. China, Institute of Information Science in 1996. He was Postdoctoral Research Associate during 2000~2003, IFP, Beckman Inst., UIUC. He worked as Research Fellow/Humboldian during 1999~2000, Chair for Pattern Recognition, University of Erlangen-Nuremberg, Erlangen, Germany. He ever was Postdoctoral Fellow/Lecturer during 1997~1999, Dept. of Computer Science & Technology, Tsinghua University of P. R. China. His research interests include facial/hand gesture/body motion tracking & modeling & recognition, static/motion segmentation, visual surveillance & monitoring, object detection & recognition, machine learning, ac-

tive vision, image-based rendering and augmented reality etc.

Thomas S. Huang Received his bachelor degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, China; and his master and Ph. D. degrees in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge, Massachusetts. He was on the Faculty of the Department of Electrical Engineering at MIT from 1963 to 1973; and on the Faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now William L. Everitt Distinguished Professor of Electlrical and Computer Engineering, and Research Professor at the Coordinated Science Laboratory, and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology and Co-Chair of the Institute's major research theme Human Computer Intelligent Interaction.

During his sabbatical leaves, Dr. Huang worked at the MIT Lincoln Laboratory, the IBM Thomas J. Watson Research Center, and the Rheinishes Landes Museum in Bonn, West Germany, and held visiting Professor positions at the Swiss Institutes of Technology in Zurich and Lausanne, University of Hannover in West Germany, INRS-Telecommunications of the University of Quebec in Montreal, Canada and University of Tokyo, Japan. He has served as a consultant to numerous industrial firms and government agencies both in the U. S. and abroad. Dr. Huang's professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He is a Member of the National Academy of Engineering; a Foreign Member of the Chinese Academies of Engineering and Sciences; and a Fellow of the International Association of Pattern Recognition, IEEE, and the Optical Society of American; and has received a Guggenheim Fellowship, an A. V. Humboldt Foundation Senior U. S. Scientist Award, and a Fellowship from the Japan Association for the Promotion of Science. He received the IEEE Signal Processing Society's Technical Achievement Award in 1987, and the Society Award in 1991. He was awarded the IEEE Third Millennium Medal in 2000. Also in 2000, he received the Honda Lifetime Achievement Award for "contributions to motion analysis". In 2001, he received the IEEE Jack S. Kilby Medal. In 2002, he received the King-Sun Fu Prize, International Association of Pattern Recognition; and the Pan Wen-Yuan Outstanding Research Award. He is a Founding Editor of the International Journal Computer Vision, Graphics, and Image Processing; and Editor of the Springer Series in Information Sciences, published by Springer Verlag. Dr. Huang initiated the first International Picture Coding Symposium in 1969, and the first International Workshop on Very Low Bitrate Video Coding in 1993. Both meetings have become regular events (held every 12～18 months), and have contributed to the research and international standardization of image and video compression. He also (together with Peter Stucki and Sandy Pentland) initiated the International Conference on Automatic Face and Gesture Recognition in 1995. This Conference has also become a regular event, and provides a forum for researchers in this important and popular field.

His current research interests include multimodal (esp. audio and visual) signal processing, analysis, and visualization, esp. in the context of Human Computer Interaction Image and Video databases: Low-level content-based image and video retrieval; relevance feedback, GUI, data mining, event detection.

基于二维关节型人体模型和 EM 算法的人体跟踪

Yu Huang Thomas S. Huang

(*Beckman Institute*, *UIUC*, *Urbana*, *Illinois*, *IL*61801, 美国)

(E-mail: {yuhuang, huang}@ifp. uiuc. edu)

摘 要 提出一种跟踪单眼图像序列中的行人,并恢复其运动参数的新方法. 在跟踪中采用了基于 SPM(Scaled Prismat Model)扩展的二维纸板人模型取代三维人体模型,以获取更快的计算速度. 作者使用 EM 算法在概率框架下进行运动估计,同时,算法也考虑了混合的运动模型和运动约束,以减小解的搜索空间. 试验结果证明了该方法的有效性.

关键词 人体跟踪,EM 算法,二维关节运动

中图分类号 TP391.41