Incremental Tracking of Human Actions from Multiple Views

Masanobu Yamamoto, Akitsugu Sato, Satoshi Kawada,
Takuya Kondo and Yoshihiko Osaki
Department of Information Engineering, Niigata University
8050 Ikarashi 2-nocho, Niigata-city, 950-2181, Japan

Abstract

This paper proposes a new method for model-based tracking of a human body in 3D motion from multiple views. The tracking is performed by estimating the pose increment of the body parts from multiple image sequences after establishing of fitting an articulated model to the human body at the initial frame. The pose increment can be obtained from solving a system of linear equations. This calculation does not depend on the number of view points. Experiments verify that the proposed method can avoid occlusion and visual degenerate from a particular view point, and track a human body in complicated motion.

1 Introduction

The measurement of human motion has played an important role in various applications in virtual reality such as performance-driven computer animation and human-computer interaction. Many systems for capturing motion have been commercially provided. Most of them needs to attach some equipments on the human body. Physically and mentally, these equipments prevent the subjects from acting freely.

The use of image sequence, which are provided from video camera, makes it easier to capture the human motion in natural style. In image sequence analysis of human actions, a use of appropriate 3D model (usually articulated model) for the body is essential, since the human body has a complicated structure and action [2, 3, 8, 11, 12].

Fitting the 3D model to the body image at each frame, it is possible to measure the 3D pose of human body in motion. Even a single camera can realize this measurement. However, there are a few problems. These are that (1) an occlusion makes it difficult to measure entire body movement, (2) a visual degenerate in motion measurement may occur when the body moves toward or away from the camera, and (3) accurate fitting model to body is difficult, since the used model is usually an approximation of real body shape. To overcome these problems, several researchers have proposed methods by using multiple camera views.

Rehg and Kanade[10] tracked hand motion using a stereo camera system. Gavrila and Davis[1] tracked dancing peoples by four cameras located at four corners of the room. Kakadiaris and Metaxas[4] used three cameras placed at front, side and top of the subject.

These existing methods involve an inherent problem in the model matching. The model matching procedure needs a large amount of computation to search images for the subject to be fitted, since the model matching is carried out in every view at every frame. The method[4] of Kakadiaris and Metaxsas selects a view which provides the most information for computational speed up. Then, the computation for search reduces into matching the model with the selected view. However, it needs computation for the view selection.

In this paper, a model matching is carried out only at the initial frame. The tracking is performed by estimating the pose increment of the body parts between two successive images. The pose increment can be obtained from solving a system of linear equations. We need not to solve so many systems of linear equations as the number of view points. The proposed method, which extends the case with a single camera, can track articulated objects (i.e. human body) observed from multiple views.

The next Section describes the theoretical aspects of our method. Experiments in Section 3 verify that the proposed method can avoid occlusion and visual degenerate from a particular view point, and obtain a reliable estimation of human action.

2 Motion Estimation from Multiple Image Sequences

This section presents a method which estimates motion parameters of human body from image sequences of multiple views.

2.1 Human Motion Model

We represent a human body by an articulated structure consisting of 12 rigid parts corresponding to head, chest, abdomen, waist, upper arms, forearms, thighs and shines, respectively. Each part of the body is approximated by a polyhedron which is made by a CAD modeler [5]. Fig.1 denotes the human body model. Each part of the body has a local coordinate system of which origin is located at a joint, and a unique label to discriminate from each other. The model has a tree structure as shown in Fig.2. The root of the tree is a abdomen, and the arrow indicates a relationship between parents and children.

Rigid motion in each part of the body is a combination of rotation and translation, denoted by a matrix Q_{b_j} and vector $\boldsymbol{S}_{b_j} = (S_{x_j}, S_{y_j}, S_{z_j})$, respectively,

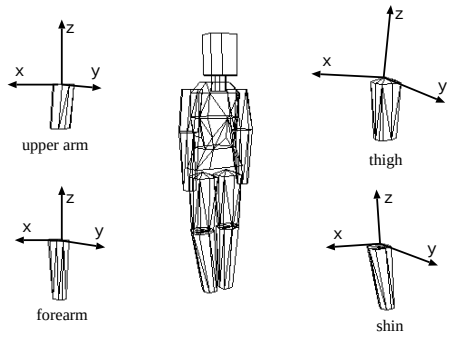

Figure 1: Body model

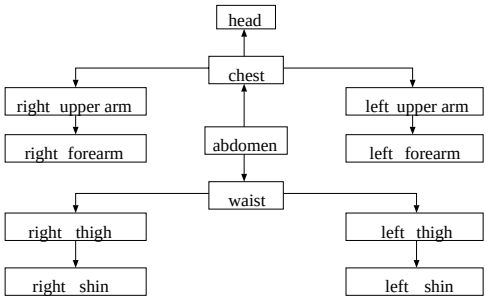

Figure 2: Articulated model of human body

where the j identifies the number of the part. When a point on the part moves from $\boldsymbol{p}_j$ to $\boldsymbol{p}'_j$, both positions are related by

$$\boldsymbol{p}'_j = Q_{b_j}\boldsymbol{p}_j + \boldsymbol{S}_{b_j} \tag{1}$$

Supposing the movement to have small rotation, the rotational matrix, Q_{b_j}, is given by

$$\mathrm{Q_{b_j}} = \begin{pmatrix} 1 & -\theta_{z_j} & \theta_{y_j} \\ \theta_{z_j} & 1 & -\theta_{x_j} \\ -\theta_{y_j} & \theta_{x_j} & 1 \end{pmatrix} \tag{2}$$

where the θ_{x_j}, θ_{y_j} and θ_{z_j} are small rotations around x, y and z axes, respectively.

We define a set of these rotational and translational quantities of all parts of the body as a vector, ϕ, of motion parameters:

$$\begin{aligned} \phi \quad = \quad & (S_{x_1}, S_{y_1}, S_{z_1}, \theta_{x_1}, \theta_{y_1}, \theta_{z_1}, \cdots, \\ & S_{x_j}, S_{y_j}, S_{z_j}, \theta_{x_j}, \theta_{y_j}, \theta_{z_j}, \cdots)^\top \end{aligned}$$

2.2 Camera Model

Let us set each camera coordinate system in (x_i, y_i, z_i), where the i denotes the camera number. Supposing a central projection, a point in the space is mapped onto an image point, (X_i, Y_i), on an image plane, $z_i = 1$, by the relation

$$\begin{cases} X_i = f_{x_i}\dfrac{x_i}{z_i} \\ Y_i = f_{y_i}\dfrac{y_i}{z_i} \end{cases} \tag{3}$$

where f_{x_i} and f_{y_i} are scale factors in the x and y direction, respectively.

A spatial displacement $(\delta x_i, \delta y_i, \delta z_i)$ and the corresponding image displacement $(\delta X_i, \delta Y_i)$ can be related by

$$\begin{cases} \delta X_i &= (f_{x_i}\delta x_i - X_i\delta z_i)/z_i \\ \delta Y_i &= (f_{y_i}\delta y_i - Y_i\delta z_i)/z_i \end{cases} \tag{4}$$

as far as these displacements could be small.

A point, $\boldsymbol{p}_{c_i}$, represented in the camera coordinate is related to the representation, $\boldsymbol{p}_w$, in the inertia coordinate by

$$\boldsymbol{p}_w = R_{c_i}\boldsymbol{p}_{c_i} + \boldsymbol{T}_{c_i} \tag{5}$$

where R_{c_i} and $\boldsymbol{T}_{c_i}$ denote the pose of the camera coordinate system with respect to the inertia coordinate.

2.3 Camera Calibration

The intrinsic and extrinsic camera parameters can be obtained from a camera calibration.

In this paper, we use a camera calibration method proposed by Lowe[7]. We locate a calibration object, of which size is known, in the scene, and suppose the object-oriented coordinate to be an inertia coordinate. The left of Fig.3 shows the object image and the model projection on the image based on appropriate camera parameters. The calibration method iteratively calculates the camera parameters so that a ridge of the model could be aligned with the corresponding edge on the image. Fitting the model to the object image gives the intrinsic parameters, f_{x_i} and f_{y_i}, and extrinsic parameters, R_{c_i} and $\boldsymbol{T}_{c_i}$.

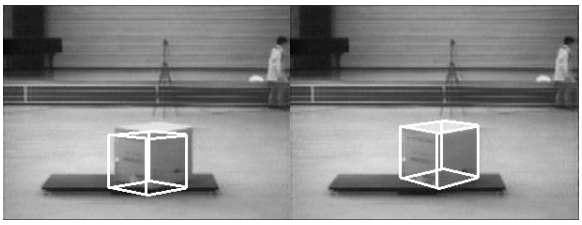

Figure 3: Before (left) and after (right) camera calibration

2.4 Model Fitting at Initial Frame

At an initial frame to start for tracking, the model fitting to human body is established by manual manner as follows.

The body model can be projected on the image plane based on the camera parameters obtained from camera calibration. We manually adjusted the pose of

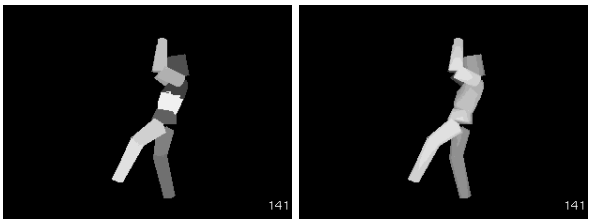

Figure 4: The left is a region image where each gray level denotes a label assigned to a part of the body, and the right is a depth image where a gray level denotes a depth from camera. Both images are examples viewed from camera 1 in Fig.10.

the model whose projection could be fitted to the body image. A transformation between the j-th object-oriented coordinate and the inertial coordinate is described by

$$\boldsymbol{p}_w = R_{b_j}\boldsymbol{p}_{b_j} + \boldsymbol{T}_{b_j} \tag{6}$$

where $\boldsymbol{p}_w$ and $\boldsymbol{p}_{b_j}$ indicate the same point represented by the inertia coordinate and the j-th object-oriented coordinate, respectively, and also R_{b_j} and $\boldsymbol{T}_{b_j}$ are the rotational and translational transformations, respectively, of the object-oriented coordinate with respect to the inertia coordinate.

Fitting the human body model to the human image gives transformations, R_{b_j} and $\boldsymbol{T}_{b_j}$. Then, the CAD modeler calculates a region image (left in Fig.4) and depth image (right in Fig.4) from a projection of the model. The region image in which pixel corresponds to each body part denotes an occlusion and appearance from the camera view. In the case with a single camera, since the body model is an approximation of real shape of the body unlike the calibration object, it is difficult to determine a 3D pose of the model. Model fitting from multiple views, however, can result into measurements of the 3D pose of human body.

2.5 Estimation of Motion Parameters

In this section, an equation will be derived so as to estimate motion parameters from image sequences of multiple views.

When the human model fits to the human image, coordinates of the point on the human body are given as coordinates of the corresponding point on the model by the depth image. Let us a movement of the point be a function of the motion parameters. Consider the simple model as shown in Fig.5, there are two bodies where bodies 1 and 2 are a parent and a child, respectively.

Suppose that a point, $\boldsymbol{p}_i$, on the body 2 moves to a new location,$\boldsymbol{p}'_i$, where both $\boldsymbol{p}_i$ and $\boldsymbol{p}'_i$ are represented by a camera coordinates systems. The new location can be calculated from coordinate transformations and body motions in following steps.

a. Transform from the camera coordinate into an inertia one.

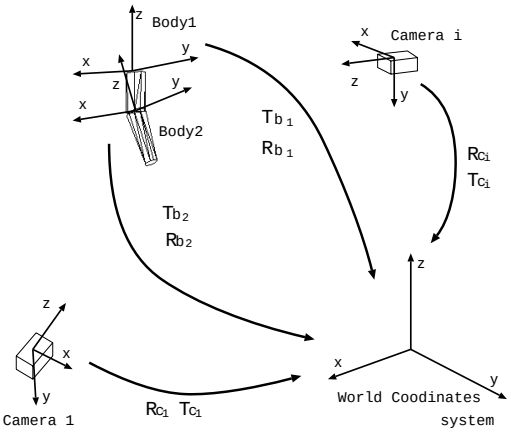

Figure 5: The camera-centered,body,and world coordinates system

b. Transform from the inertia coordinate into an object-oriented one.

c. Move the point with respect to the object-oriented coordinate.

d. Transform from the object-oriented coordinate into the inertia one.

e. Transform from the inertia coordinate into a parent-oriented one.

f. Move the point with respect to the parent-oriented coordinate.

g. Transform the parent-oriented coordinate into the inertia one.

h. Transform from the inertia coordinate into the camera one.

These steps results into

$$\begin{aligned}\boldsymbol{p}'_i &= R_{c_i}\left(R_{b_1}\left(Q_{b_1}R_{b_1}^{-1}\left(R_{b_2}\left(Q_{b_2}R_{b_2}^{-1}\right.\right.\right.\right.\\ &\quad \left.\left.\left.\left.(R_{c_i}\boldsymbol{p}_i + \boldsymbol{T}_{c_i} - \boldsymbol{T}_{b_2}) + \boldsymbol{S}_{b_2}\right)\right.\right.\right.\\ &\quad \left.\left.\left. + \boldsymbol{T}_{b_2} - \boldsymbol{T}_{b_1}\right) + \boldsymbol{S}_{b_1}\right) + \boldsymbol{T}_{b_1} - \boldsymbol{T}_{c_i}\Big)\end{aligned} \tag{7}$$

If the body 1 has more parent, repeat steps (e), (f) and (g) to calculate the effect of the parent motion.

The $\boldsymbol{p}'_i$ is a function with only motion parameters, since the camera parameters and the initial pose of body parts can be given by methods described in previous Sections.

Expanding $\boldsymbol{p}'_i$ about zero values and neglecting second and higher order terms, 3-D displacement, $\delta\boldsymbol{p}_i = \boldsymbol{p}'_i - \boldsymbol{p}_i$, is represented by a linear combination of motion parameters as

$$\delta\boldsymbol{p}_i = \boldsymbol{p}'_i - \boldsymbol{p}_i$$

$$
\begin{aligned}
&= (\delta x_i, \delta y_i, \delta z_i)^\top \\
&= \sum_{j=1}^{2} \left\{ \frac{\partial \boldsymbol{p}'_i}{\partial S_{x_j}} S_{x_j} + \frac{\partial \boldsymbol{p}'_i}{\partial S_{y_j}} S_{y_j} + \frac{\partial \boldsymbol{p}'_i}{\partial S_{z_j}} S_{z_j} \right. \\
&\quad \left. + \frac{\partial \boldsymbol{p}'_i}{\partial \theta_{x_j}} \theta_{x_j} + \frac{\partial \boldsymbol{p}'_i}{\partial \theta_{y_j}} \theta_{y_j} + \frac{\partial \boldsymbol{p}'_i}{\partial \theta_{z_j}} \theta_{z_j} \right\} \\
&= J_i(\phi)\phi \qquad (8)
\end{aligned}
$$

where $J_i(\phi)$ is the Jacobian matrix.

Let $E(X_i, Y_i, t)$ be an image sequence from the i-th camera. The image displacement, $(\delta X_i, \delta Y_i)$, at point, (X_i, Y_i), after a unit time period is constrained by the equation

$$E_{X_i}\delta X_i + E_{Y_i}\delta Y_i + E_t = 0 \qquad (9)$$

where (E_{Xi},E_{Yi}) and E_t are spatial and temporal gradients, respectively.

Substituting δx_i, δy_i and δz_i of eq.(8) into eq.(4), and δX_i and δY_i of eq.(4) into eq.(9), a linear equation with motion parameters, ϕ, as unknowns is derived as follows.

$$E_{pi}J_i(\phi)\phi = E_t \qquad (10)$$

where

$$E_{pi} = \left[-\frac{f_{x_i}E_{Xi}}{z_i}, -\frac{f_{y_i}E_{Yi}}{z_i}, \frac{X_iE_{X_i} + Y_iE_{Y_i}}{z_i} \right]$$

and depth z_i is given from the depth image when the human model aligns with the human body.

2.6 Estimation of Motion Parameters from Multiple Views

We can determine which part of the body a point on the image taken by the i-th camera belongs to by using the region image. We get a system of linear equations which correspond to these points.

$$A_i\phi = B_i \qquad (11)$$

where A_i and B_i are the coefficient matrix and a constant vector, respectively.

Supposing the number of cameras to be n, we have the n systems of linear equations which correspond to cameras. Since the unknown ϕ belongs to the unique model of the human body, we can get one large system of linear equations by gathering these n systems.

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{bmatrix} \phi = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{bmatrix} \qquad (12)$$

The motion parameters, ϕ, can be obtained from a least square solution of eq.(12). The estimation of ϕ needs not to solve eq.(11) in camera by camera, but solve eq.(12) only once. Therefore, this computation does not depend on the number of cameras.

Moving the model based on the estimated motion parameters, the moved model also fits to the image of the human body at the next frame. Repeating this procedure in frame by frame enables the computer to track the human body in motion.

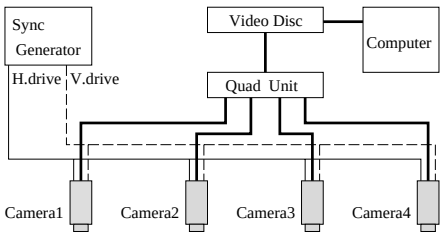

Figure 6: A system for acquisition of multiple image sequences

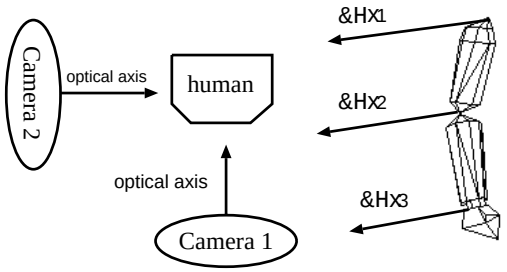

Figure 7: The left figure is a top view of an arrangement of two cameras, and the right figure is a CAD model of the human arm.

3 Experiments on Advantages of Multiple Views

We will give following some experiments to verify several advantages brought by using multiple views.

a. Avoiding visual degenerate: Vanishing an ambiguity of motion estimation in the direction of the optical axis from a single view.

b. Tracking body in very complicated motion.

3.1 A System to Acquire Multiple Image Sequences

We show a system to acquire image sequences from four view points. Fig.6 denotes this system.

All four cameras are driven by the same external synchronous signal. Four video images, which are completely synchronized, are transferred to the quad unit (SONY, YS-Q400). The unit reduces each full-size image to a quarter-size image, and provides a new full-size image by gathering the four quarter-size images. The output image of the quad unit is recorded in the optical video disc recorder (Panasonic, LQ-4100).

The number of cameras, 4 here, is limited by the capacity of the quad unit system. The more systems will be available for more cameras.

3.2 Avoiding Visual Degenerate

When an object moves towards or away from a camera, it is difficult to estimate the movement with high reliability, since the apparent movement is generally

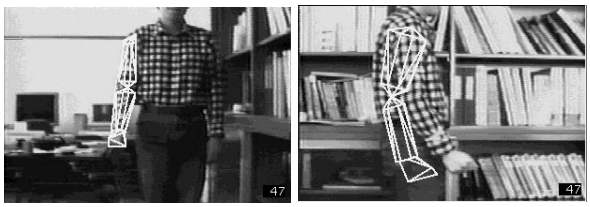

Figure 8: Tracking result from a view of Camera 1 alone

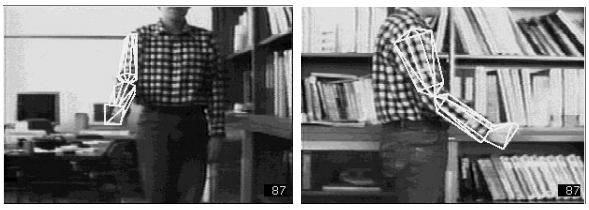

Figure 9: Tracking result from multiple views of Cameras 1 and 2

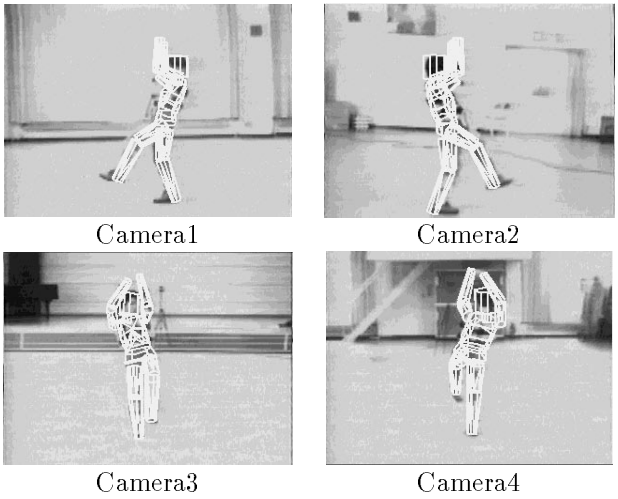

Figure 10: Fitting a model to the body images at initial frames

very small. In fact, we can show the difficulty by an experiment as shown in the left of Fig.7, where the camera 1 facing to the human body observes that the subject is lifting his right hand from the bottom to the front.

With the sake of simplicity, the model is composed with an upper arm and a forearm, and each part rotates around only one axis such as the right of Fig.7. The left image of Fig.8 denotes tracking result by only camera 1. The tracking result can be checked by another camera (camera 2 in the left of Fig.7) which observes the side view of the subject. The right image of Fig.8 shows a separation between the model and image of the arm at one frame during tracking. In Fig.9, we denotes the tracking result by simultaneous use of the both cameras. Observations from the both cameras shows successful tracking result.

Since the camera 2 can avoid the degenerate from view of camera 1, it may seem that the use of the both camera views is meaningless. However, it is important that there is no need to select a camera which can give the most reliable estimation.

3.3 Very Complicated Motion Tracking

As one example of very complicated motion, we try to track a pitching motion in the baseball game.

In this experiment, four cameras are located around the pitcher. The model used here is composed of 12 parts of body as shown in Fig.1, where the pitching motion is represented by 39 motion parameters, that is, three degrees of rotational freedom to all parts and six degrees of freedom to the abdomen.

Fig.10 shows the model fitting at initial frame. Fig.11 shows a tracking result viewed from the camera 3 at intervals of 2 frames.

In the Fig.10, we can forecast some difficulties in tracking by a single camera. In fact, the right arm of the subject is occluded in the view of camera 1. Also, the left arm is not appeared in the view of camera 2. Moreover, the camera 3 cannot clearly capture the left leg's motion along the optical axis. The proposed method can estimate all motion parameters by solving a large system of linear equations derived from four cameras.

4 Conclusion

This paper proposed a model-based method for visual tracking of human body from multiple views. This tracking is performed by integrating pose increments onto the initial pose. The calculation for the pose increment does not depend on the number of view points. Experiments verify that the proposed method can avoid occlusion and visual degenerate from a particular view point, and track a human body in pitching.

Since this tracking method can share the human body model together with computer graphics of creating human figures, it is possible to build a compact system for performance-driven animation. We will show an animation based on this system.

The proposed method may fail in the tracking during a long sequence of images, since error in tracking is also accumulated even the small error. Causes for the error are an inaccuracy of the camera calibration, an approximation in the gradient-based equation (9), incoherent movements at occlusion boundaries and the geometrically rough model of human body. A length of successful tracking depends on the speed and complexity of motion and texture patterns on the body and background. Now, the most successful experiment has 500 frames.

As for future works, we have to pursue automatic model fitting at the initial frame. One idea is to com-

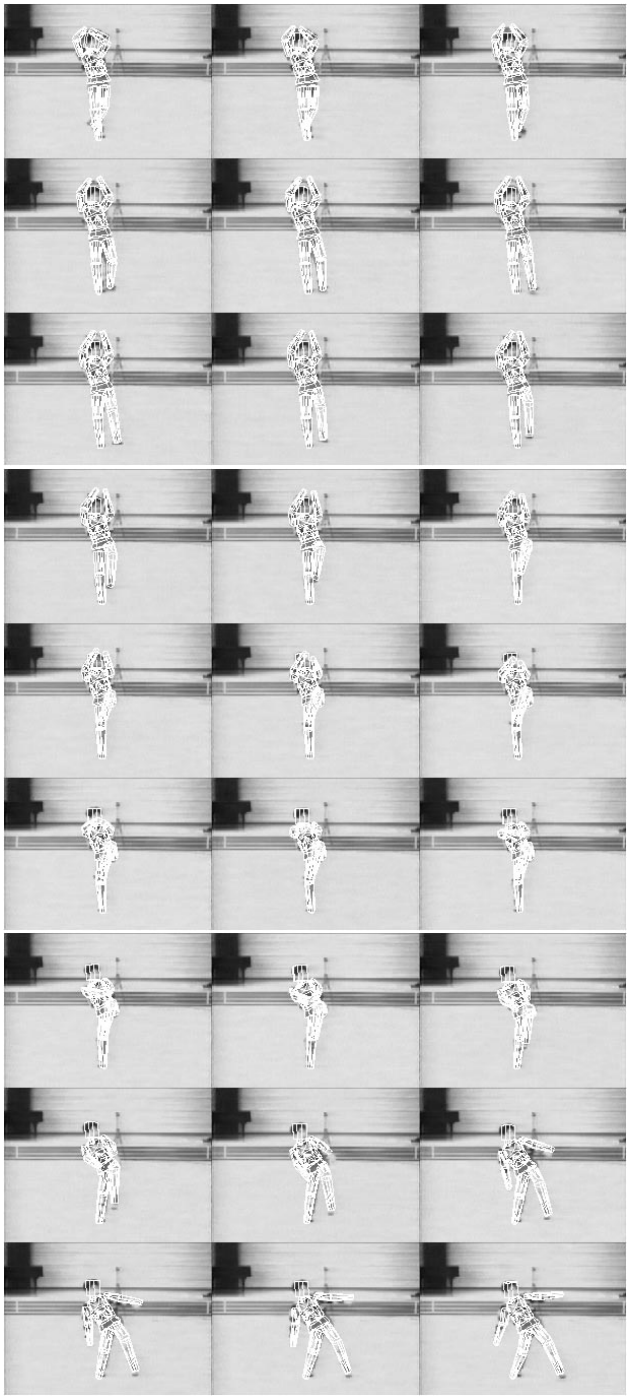

Figure 11: Tracking result of pitcher's motion

bine the proposed method with existing model fitting techniques [1, 4, 9, 10]. Also, this compromise will be able to correct a separation of model and body image during tracking in long time by model fitting at appropriate key-frames. An alternative is fitting the 3D model to 3D coordinates measured by stereo camera system.

References

[1] D.M.Gavrila and L.S.Davis: 3-D model-based tracking of humans in action: a multi-view approach, Proc. of IEEE CVPR'96, pp.73–80, 1996.

[2] L.Goncalves, E.D.Bernardo, E.Ursella and P.Perona: Monocular tracking of the human arm in 3D, Proc. of 5th ICCV, pp.764–770, 1995.

[3] D.Hogg: Model-based vision: a program to see a walking person, *Image and Vision Computing*, Vol.1, No.1, pp.5-20, 1983.

[4] I.A.Kakadiaris and D.Metaxas: Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection, Proc. of IEEE CVPR'96, pp.81–87, 1996.

[5] K.Koshikawa and Y.Shirai: A 3-D modeler vision research. In *Proc. of the Int. Conference on Advanced Robotics*, pp.185-190, 1985.

[6] M.K.Leung and Yee-Hong Yang: A region based approach for human body motion analysis, *Pattern Recognition*, Vol.20, No.3, pp.321-339, 1987.

[7] D.G.Lowe: Fitting parameterized three dimensional models to images , IEEE PAMI , Vol.13 , No.5 , pp.441-450 , 1991.

[8] J.O'Rourke and N.J.Badler: Model-based image analysis of human motion using constraint propagation, IEEE PAMI,Vol.2, No.6, pp.522-536, 1980.

[9] J.Ohya and F.Kisino: Human posture estimation from multiple images using genetic algorithm, 12th ICPR, pp.750-753, 1994.

[10] J.M.Rehg and T.Kanade: DigitEyes: Vision-based human hand tracking, CMU-CS-93-220, 1993.

[11] K.Rohr: Towards model-based recognition of human movements in image sequences, *CVGIP: Image Understanding*, Vol.59, No.1, pp.94-115, 1994.

[12] M.Yamamoto and K.Koshikawa: Human motion analysis based on a robot arm model, Proc. of CVPR'91, pp.664–665, 1991