Marker-less Human Motion Estimation using Articulated Deformable Model

Koichi Ogawara and Xiaolu Li and Katsushi Ikeuchi

***Abstract*— This paper presents a novel whole body motion estimation method by fitting a deformable articulated model of the human body into the 3D reconstructed volume obtained from multiple video streams. The advantage of the proposed method is two fold: (1) combination of a robust estimator and ICP algorithm with Kd-tree search in pose and normal space make it possible to track complex and dynamic motion robustly against noise and interference between limb and torso, (2) the hierarchical estimation and backtrack re-estimation algorithm enable accurate estimation.**

The power to track challenging whole body motion in real environment is also presented.

I. INTRODUCTION

Motion capture technique is used for measuring whole body motion in various applications: including making character animation in film and game industries, biomechanical analysis and engineering on ergonomics and human factors. Demand for whole body motion is also increasing in the field of robotics, such as gesture based user interface and motion generation of a humanoid robot[1].

However, the state-of-the-art commercial systems, for example an optical or magnetic motion capture system, are expensive and generally force a user to put on restrictive markers all over the body.

In this paper, a novel marker-less motion capture technique is proposed that can track complex and dynamic whole body motion robustly from multiple video streams.

A. Previous work

There is a growing demand for measuring whole body motion in daily environment for surveillance, user interface and so on. For this purpose, easy-to-use and unconstrained marker-less motion capture technique has been an active research topic for the last decade[2].

In this technique, a user is observed by a single or multiple video cameras and captured images are processed to estimate the user's pose using computer vision techniques. Single camera approach[3], [4], [5] is convenient, however occlusion and ambiguity is difficult to be solved in a monocular framework, thus this approach is appropriate for rough pose estimation used in gesture and motion pattern recognition.

Multiple cameras approach is typically used to estimate the accurate motion[6], [7], [8], [9], [10] and it gains much attention in the hope that it will replace the state-of-the-art motion capture systems.

K. Ogawara is with Faculty of Engineering, Kyushu University, Fukuoka, JAPAN ogawara@is.kyushu-u.ac.jp

X. Li is with the University of Tokyo, Tokyo, JAPAN

K. Ikeuchi is with faculty of Interfaculty Initiative in Information Studies, the University of Tokyo, Tokyo, JAPAN

The latter approach can be categorized into 2 major classes. The first class is called motion tracking. Suppose the pose at time $t-1$ is known, the pose at time t is expected to be within the neighborhood of the pose at time $t-1$. Thus, the pose at time t is efficiently estimated by searching for the local minimum of some error functions only around the pose at time $t-1$. Delamarre et al. defined an articulated body model based on cylinder and rectangular parallelepiped, and sequentially estimated the latest pose based on the previous estimation by minimizing the distance between the silhouette in the captured images and the edges of the body model projected onto the same images using a gradient descent method[7].

However, this class needs initial solution and also needs to estimate the pose from a scratch when tracking fails. The second class is to infer the pose directly from images using some exemplar-based learning techniques. Gavrila et al. defined a kinematic model of the human body by using super quadrics, and built a data base that learns numerous pairs of 2D projection of the body model and its corresponding pose parameters. Input images from 4 cameras are compared with the images in the data base using chamfer distance and the pose parameters corresponding to the most similar one is retrieved[6].

What captured onto images is a surface of the body, thus some researchers have proposed to explicitly handle the deformable skin in their body models for accurate pose estimation. Illic et al. represented a skin model by using implicit functions and estimated the shape of the upper body using a Free Form Deformation (FFD) method[9]. However, this model doesn't have joint structure inside, so it is not appropriate for motion estimation.

Cheung et al. proposed a method to estimate both the structure of the articulated body model, i.e. joint position, and the motion simultaneously. They divided colored surface points (CSP) into groups of different rigid motion and estimated the rigid transformation for each group in consecutive images[8]. CSP is a 3D point on the body surface confirmed using multiple-view geometry. However, generating an articulated body model is a complex process and it is not always possible to estimate all the joint positions correctly.

Kehl et al. proposed a method to estimate the motion by using a deformable skin model with joint structure[10]. The 3D shape of the user is reconstructed using a volume intersection method and the points on the volume are searched for the nearest one to each vertex on the skin model. These correspondences are used to move the pose and the joints to the right direction by minimizing the sum of errors using a

gradient descent method.

In this paper, we assume continuity in motion and deal with the problem of estimating the pose at time t when the pose at time $t-1$ is given. As for the pose at time 0, any exemplar-based approach such as[5] can be used.

Our method is most similar to Kehl's approach. We use a similar deformable model with joint structure against the 3D reconstruction from multiple video streams. The advantage of our method is two fold:

1) Combination of a robust estimator and ICP with Kd-tree search in pose and normal space make it possible to track complex and dynamic motion robustly against noise and interference between limb and torso
2) The hierarchical estimation and backtrack re-estimation process enable accurate estimation

In the following, the body model composed of a deformable skin and joint structure is described in Section 2. In Section 3, the proposed motion estimation method is explained in detail. Our method has been verified in real environment and the results are evaluated in Section 4. Finally, we conclude in Section 5.

II. Articulated Deformable Model

A. Definition of the body model

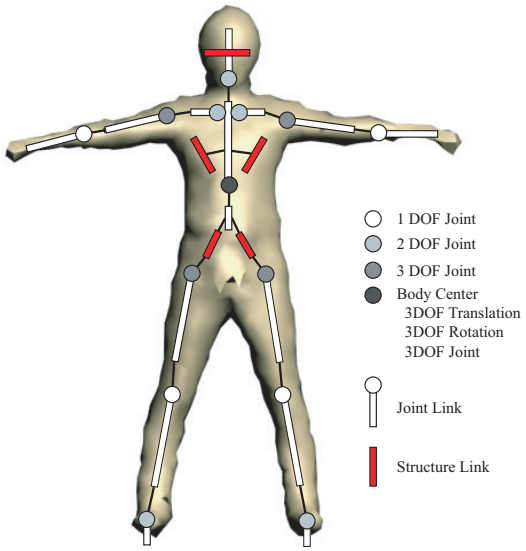


Fig. 1. Body model

The body model used in this research is composed of the link model that represents the joint structure and the skin model that represents the surface of the body as shown in Fig.1. The skin model is deformed naturally following the change in the joint angles.

The link model has 29 DOF for joints, 3 DOF for body translation and 3 DOF for body rotation. Since the position of each vertex on the skin model is affected by the links in the neighborhood as described later, we introduce two types of links in the link model. The first type is a joint link that connects two joints in parent-child relationship. The other is a structure link that has no joint but adds partial stiffness to the skin model like a rib to avoid unnatural deformation of the skin model.

The skin model is obtained by the 3D reconstruction process explained in Section IV-A.

B. Skin deformation

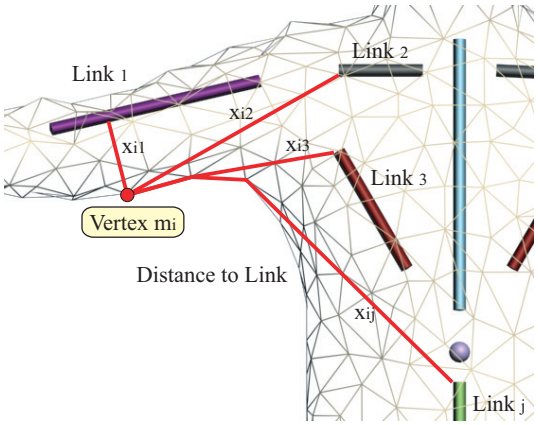


Fig. 2. Distance between a vertex on the skin model and a link

In the surface deformation method[11] used in Kehl's approach[10], control points are assigned along the link structure. Their positions are moved or they degenerate according to the change in the joint angles, and the smooth surface is generated from these control points. However, it cannot handle the situation where several links which have no direct connection to each other gather close under the skin model as is the case with the shoulder in Fig.2.

To solve this problem, the position of i-th vertex on the skin model in the body centered coordinates frame is determined to be $m'_i(\boldsymbol{\theta})$, a function of the joint angles $\boldsymbol{\theta}$.

As shown in Fig.2, the distance $x_{i,j}$ between the i-the vertex and the j-th link is defined as the length of the shortest path from the i-th vertex to the j-th link under the constraint that the path never penetrate the skin model. To calculate $x_{i,j}$, a graph is constructed where all the vertices in the skin model and all the links are defined as nodes, and all the edges in the skin model are defined as arcs. Then, for each vertex in the skin model, an arc from the vertex to j-th link is added to the graph if a segment from the vertex to j-th link doesn't intersect with the skin model. The weight of an arc is set as the Euclid distance between its terminal nodes. Finally, the distance $x_{i,j}$ is calculated using Dijkstra's shortest path algorithm.

Next, the weight between the i-th vertex and the j-th link is defined by a function $w(x_{i,j})$ as in Eq.(1), and the new position of the i-th vertex is defined as a sum of weighted positions as in Eq.(2).

$$w(x_{ij}) = a \cdot e^{-bx_{ij}} \qquad \sum_j w(x_{ij}) = 1 \tag{1}$$

$$\begin{pmatrix} \boldsymbol{m}'_i(\boldsymbol{\theta}) \\ 1 \end{pmatrix} = \sum_{j=1}^{L} w(x_{ij}) \cdot \boldsymbol{T}_j(\boldsymbol{\theta}) \cdot \boldsymbol{T}_j^{-1}(\boldsymbol{0}) \cdot \begin{pmatrix} \boldsymbol{m}_i \\ 1 \end{pmatrix} \tag{2}$$

where a and b in Eq.(1) are constant determined empirically. $T_j(\boldsymbol{\theta})$ is a transformation matrix from the j-th link's coordinates frame to the body centered coordinates frame given the joint angles $\boldsymbol{\theta}$. m_i is the position of the i-the vertex in the body centered coordinates frame when all the joint angles equal to 0, that is exactly the pose shown in Fig.1.

III. Motion Estimation

A. Formulation

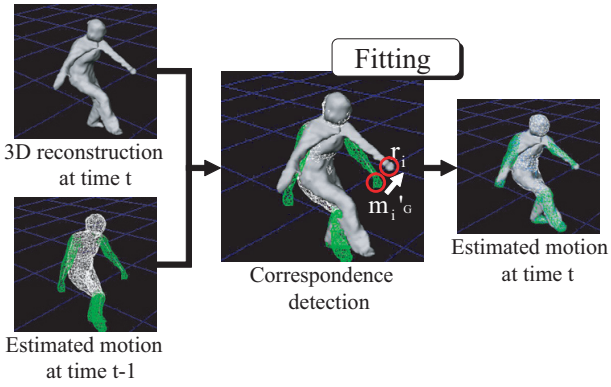


Fig. 3. Motion estimation based on correspondence detection

By applying a method explained in Section IV-A to multiple video streams, 3D shape of the target user at time t can be reconstructed. If the motion parameters, i.e. the joint angles $\boldsymbol{\theta}$, body translation $\boldsymbol{t}$ and body rotation $\boldsymbol{R}$, at time $t-1$ are given, the motion estimation problem is formulated as a minimization problem of Eq.(3) given the motion parameters at the previous time frame.

$$E = min \sum_i \parallel m_i'^G - r_i \parallel^2 \tag{3}$$

where $m_i'^G$ is the position of the i-th vertex in the skin model and r_i is the corresponding point on the reconstructed volume as show in Fig.3.

This formulation is an extension of the ICP algorithm[12], a rigid body alignment algorithm, to a deformable object alignment algorithm. We have already proposed the basis of this formulation in the hand shape estimation problem[13]. In this paper, several new algorithms are introduced to enhance the original algorithm for tracking much complex motion including whole body motion.

If the error in Eq.(3) follows a gaussian distribution, the motion parameters can be estimated by minimizing Eq.(3) by solving the least squares method.

But the real error distribution usually doesn't follow a gaussian distribution because of measurement errors and occlusion, thus the effect of outliers makes the localization process unstable. Therefore, Wheeler proposed a technique to apply M-estimator to approximate the real error distribution[14]. M-estimator is a generalized form of the least squares method and is formulated as Eq.(4).

$$E = min \sum_i \rho(\parallel \boldsymbol{R} \cdot m_i'(\boldsymbol{\theta}) + \boldsymbol{t} - r_i \parallel^2) \tag{4}$$

where $\rho(z)$ is a function of the error z.

The motion parameters $\boldsymbol{p} = (\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{R})$ that satisfy Eq.(4) are calculated by solving Eq.(5) equals 0.

$$\frac{\delta E}{\delta \boldsymbol{p}} = \sum_i \frac{\delta \rho(z_i)}{\delta z_i} \frac{\delta z_i}{\delta \boldsymbol{p}} \tag{5}$$

where $z_i = \parallel \boldsymbol{R} \cdot m_i'(\boldsymbol{\theta}) + \boldsymbol{t} - r_i \parallel^2$.

Here, we introduce a weight function $e(z)$ that represents errors as in Eq.(6).

$$e(z) = \frac{1}{z} \frac{\delta \rho}{\delta z} \tag{6}$$

Then, Eq.(5) can be rewritten as Eq.(7). If we ignore the fact that $e(z)$ is a function of z, this is a form of weighted least squares.

$$\frac{\delta E}{\delta \boldsymbol{p}} = \sum_i e(z_i) z_i \frac{\delta z_i}{\delta \boldsymbol{p}} \tag{7}$$

In this study, Lorentzian distribution is chosen as a probability distribution of errors to exclude the effect of outliers and the weight function $e(z)$ is defined as in Eq.(8).

$$e(z) = \left(1 + \frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right)^{-1} \tag{8}$$

Eq.(4) can be solved using the conjugate gradient method and $\boldsymbol{p}$ that minimizes the error is obtained.

B. Correspondence detection by Kd-tree search in pose and normal space

To solve Eq.(4), we have to find r_i on the reconstructed volume corresponding to m_i'. If we simply choose the nearest point in Euclidean 3-space, many false correspondences can be detected especially in the case of an articulated object as shown in Fig.4.

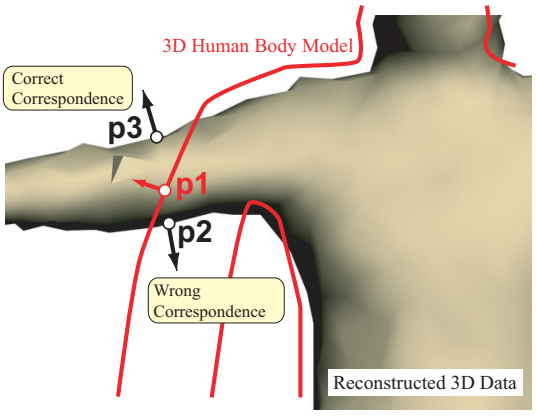


Fig. 4. Correspondence detection

To avoid this problem, the similarity in the normal vector is also considered and correspondence detection is performed both in pose space and normal vector space under Kd-tree framework.

In Euclidean 3-space, Kd-tree is built by recursively partitioning the data points into two groups along one of

x,y,z axes where the variance is maximum. In this case, p_2 becomes the corresponding point to p_1 as shown in Fig.4.

In our method, variance of the normal vector is defined as in Eq.(9).

$$\text{normal\_variance} = w \cdot \arccos(\boldsymbol{n}_i \cdot \boldsymbol{n}_j)^2 \qquad (9)$$

where $\boldsymbol{n}_i$ is the normal vector of the i-th vertex. w is determined to be the squared scale of the bounding box of the skin model so as to balance the variance in Euclidean 3-space.

Data points are sequentially partitioned along one of x,y,z, normal axes where the variance is maximum. In this case, p_3 becomes the corresponding point to p_1 successfully as shown in Fig.4.

C. Hierarchical estimation

When minimizing Eq.(4), if we try to estimate all the motion parameters $\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{R}$ simultaneously, the terminal links in the link structure tend to fall into local minimum and it prevents the other links from being aligned correctly.

Therefore, the hierarchical estimation approach is used in that the translation and rotation of the body center, $\boldsymbol{t}, \boldsymbol{R}$, are estimated first and then the joint angles, $\boldsymbol{\theta}$, are estimated in order from those around the body center, $\boldsymbol{\theta_1}$, to their descendants, $\boldsymbol{\theta_n}$, as shown in Fig.5.

To minimize the effect from the not-yet-aligned parts in the body model, only the vertices on the skin model around the joint of interest are used during each step of the hierarchical estimation.

repeat

1.Store the old motion parameters
$(\boldsymbol{t}, \boldsymbol{R}) = (\boldsymbol{t'}, \boldsymbol{R'})$
$(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \cdots, \boldsymbol{\theta_N}) = (\boldsymbol{\theta_1}', \boldsymbol{\theta_2}', \cdots, \boldsymbol{\theta_N}')$

2.Estimate the translation and rotation parameters
Solve $(\boldsymbol{t'}, \boldsymbol{R'})$ that minimize $E(\boldsymbol{t'}, \boldsymbol{R'}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \cdots, \boldsymbol{\theta_N})$
from the gradient $\frac{\partial E(\boldsymbol{t}, \boldsymbol{R}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \cdots, \boldsymbol{\theta_N})}{\partial(\boldsymbol{t}, \boldsymbol{R})}$

3.Estimate the first joint angle group in the hierarchy
Solve $\boldsymbol{\theta_1}'$ that minimize $E(\boldsymbol{t'}, \boldsymbol{R'}, \boldsymbol{\theta_1}', \boldsymbol{\theta_2}, \cdots, \boldsymbol{\theta_N})$
from the gradient $\frac{\partial E(\boldsymbol{t'}, \boldsymbol{R'}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \cdots, \boldsymbol{\theta_N})}{\partial \boldsymbol{\theta_1}}$
...

4.Estimate the N-th joint angle group in the hierarchy
Solve $\boldsymbol{\theta_N}'$ that minimize $E(\boldsymbol{t'}, \boldsymbol{R'}, \boldsymbol{\theta_1}', \boldsymbol{\theta_2}', \cdots, \boldsymbol{\theta_N}')$
from the gradient $\frac{\partial E(\boldsymbol{t'}, \boldsymbol{R'}, \boldsymbol{\theta_1}', \boldsymbol{\theta_2}', \cdots, \boldsymbol{\theta_N})}{\partial \boldsymbol{\theta_N}}$

until
$|E(\boldsymbol{t'}, \boldsymbol{R'}, \boldsymbol{\theta_1}', \boldsymbol{\theta_2}', \cdots, \boldsymbol{\theta_N}') - E(\boldsymbol{t}, \boldsymbol{R}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \cdots, \boldsymbol{\theta_N})| < \varepsilon$

Fig. 5. Order of the hierarchical estimation

Fig.6 shows that the body model is successively converged to the 3D reconstructed volume during one cycle of the hierarchical estimation.

D. Backtrack re-estimation

The estimated body model and the 3D reconstructed volume doesn't usually coincide. The primary reason is that

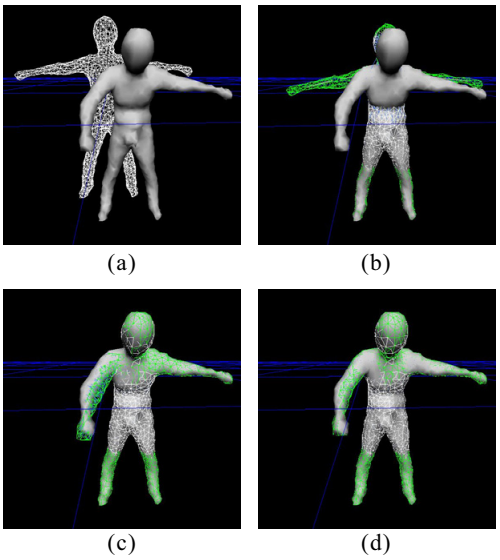


Fig. 6. Hierarchical estimation: (a) Initial state (b) The translation and rotation are estimated (c) The first joint angle group is estimated (d) The n-th joint angle group is estimated

the 3D reconstructed volume is a visual hull, a conservative estimate of the true shape, thus it gains extra volume especially around the concave parts. Another reason is that the deformation of the cloth is not considered in the body model, which adds extra dissimilarity.

Therefore, error will be accumulated during the hierarchical estimation process and a gap between the vertices on the skin model around the terminal link and the 3D reconstructed volume becomes sometimes large.

To solve this problem, we propose a method in that the residual errors backtrack from the terminal and all the joint angles in the limb are re-estimated after the hierarchical estimation. As shown in Fig.7, a local limb model that has a limited link structure corresponding to that limb and the skin model only around the terminal link is temporarily built and is localized against the 3D reconstructed volume. The residual is minimized using the same framework described in this Section, while estimating all the joint angles simultaneously.

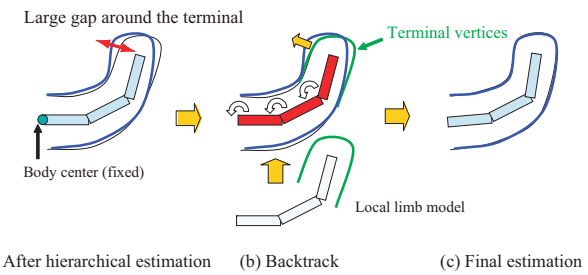


Fig. 7. Backtrack re-estimation

IV. Experimental results

A. 3D reconstruction of human motion

We setup 8 cameras, Sony DXC-9000, on the ceiling and reconstruct the 3D volume of a target from multiple video streams following the procedure below.

1) Synchronized 8 images from the cameras are stored in 30fps as shown in Fig.8(a).
2) Silhouette is extracted from each image by a background subtraction method[15] as shown in Fig.8(b).
3) The volume intersection method[16] is applied to reconstruct a 3D volume and then the Marching Cubes algorithm[17] is applied to obtain triangular mesh representation as can be seen in Fig.8(c).

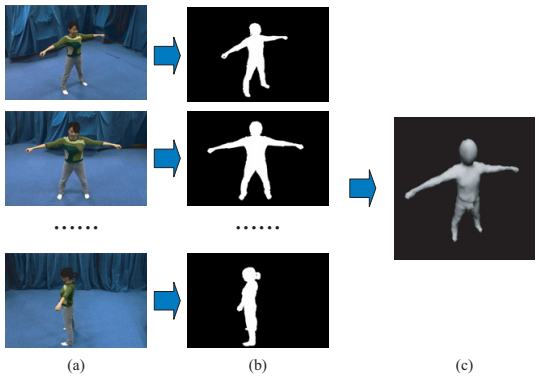


Fig. 8. 3D reconstruction of human motion

B. Evaluation

1) Evaluation of Kd-tree search in pose and normal space: Fig.9(a) shows a typical situation when Kd-tree search in pose only space is used, in that both arms adhere to the trunk. Once a limb is out of tracking and adheres to the trunk, it is difficult to judge if this pose is right or wrong automatically. Thus it is important to avoid these situations in the first place. Kd-tree search in pose and normal space is a powerful tool for this purpose and it can robustly follow the motion correctly as shown in Fig.9(b).

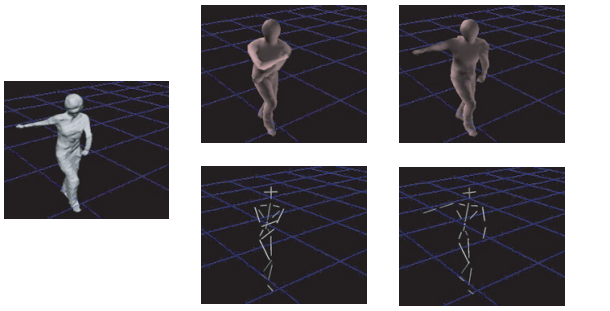


Fig. 9. Comparison between two types of Kd-tree search algorithms

2) Evaluation of the hierarchical estimation: Fig.10 shows a typical situation where the hierarchical estimation works well while the simultaneous estimation doesn't. As can be seen in Fig.10(a), the estimation process get stuck around a local minimum if all the motion parameters are estimated simultaneously.

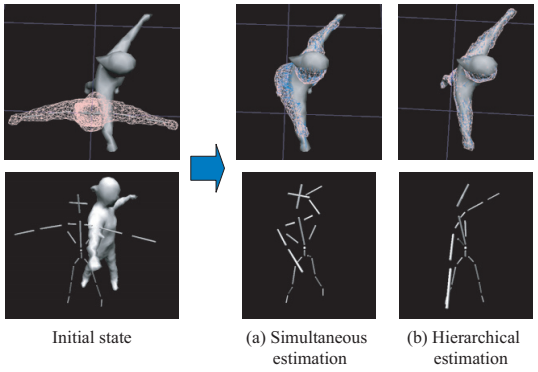


Fig. 10. Comparison between two types of estimation processes

3) Evaluation of the backtrack re-estimation: Fig.11 shows the snapshots from the result of motion estimation. The input is a series of dynamic whole body motion including jumps. The length is 400 frames, 13.3 seconds.

Fig.12 shows the plot of the mean error between the vertex on the skin model and the corresponding point on the 3D reconstructed volume for each frame. The solid line is the error from the forward with backtrack algorithm. The dotted line is the error from the forward without backtrack algorithm. Table I shows the mean and the standard deviation of each plot. They show that the algorithm with backtrack has lower mean error and is much stable.

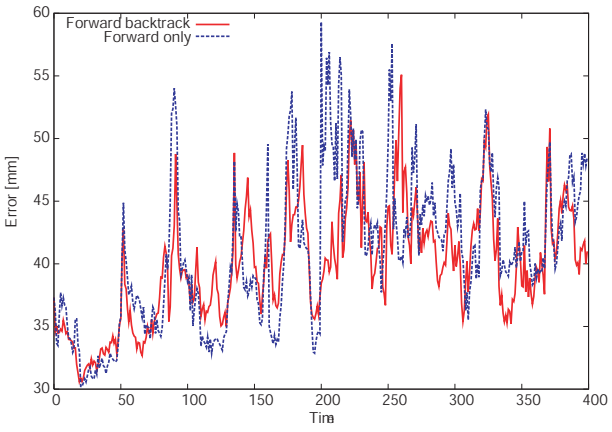
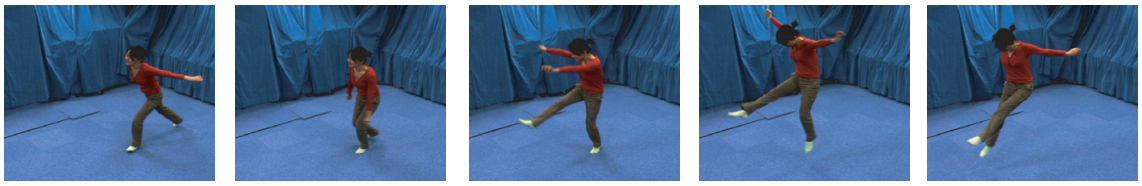


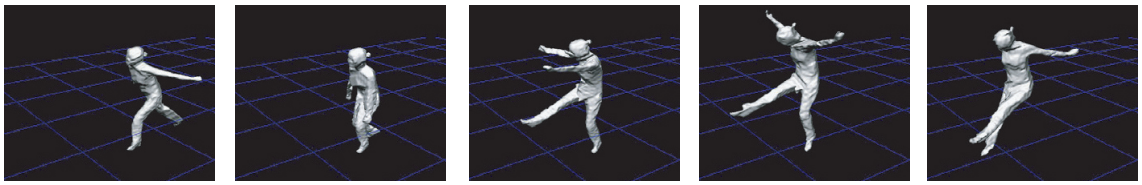
Fig. 12. Mean error in the motion estimation

V. Conclusion

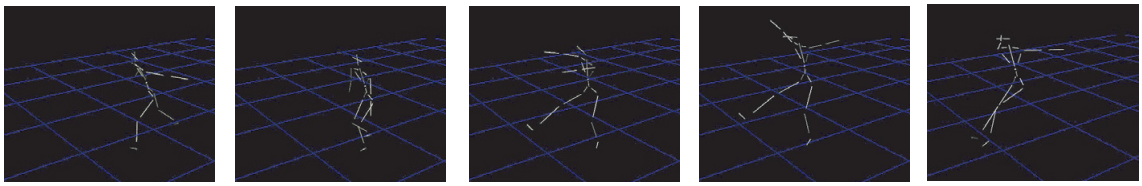
A marker-less motion capture algorithm for whole body motion is presented in this paper. Combination of a robust estimator and ICP with Kd-tree search in pose and normal space make it possible to track complex and dynamic motion robustly against noise and interference between limb and



(a) Captured images from one of the 8 cameras



(b) Reconstructed volume from capture imagess



(c) Estimated motion

Fig. 11. Result of motion estimation

TABLE I

MEAN AND STANDARD DEVIATION OF THE MOTION ESTIMATION ERROR

	Mean error [mm]	Stand. deviation [mm]
Forward with backtrack	39.899281	4.618398
Forward without backtrack	40.941326	6.035067

torso. Also, the hierarchical estimation and backtrack re-estimation process enable accurate estimation.

Future work includes the initial pose estimation, automatic model scale adjustment for tracking people of various height and generation of humanoid robot motion.

ACKNOWLEDGMENTS

This work is supported by the Japan Science and Technology Agency (JST) under the CREST "Foundation of technology supporting the creation of digital media contents" project.

REFERENCES

[1] S. Nakaoka, A. Nakazawa, K. Yokoi, and K. Ikeuchi, "Leg motion primitives for a dancing humanoid robot," in *Int. conf. on Robotics and Automation*, 2004, pp. 610–615.

[2] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding: CVIU*, vol. 81, no. 3, pp. 231–268, 2001. [Online]. Available: citeseer.ist.psu.edu/moeslund01survey.html

[3] J. M. Rehg and T. Kanade, "Model-based tracking of self-occluding articulated objects," in *IEEE International Conference on Computer Vision: ICCV*, 1995, pp. 612–617. [Online]. Available: citeseer.nj.nec.com/201852.html

[4] Y. K. N. Shimada, Y. Shirai and J. Miura, "3-d pose estimation and model refinement of an articulated object from a monocular image sequence," in *The 3rd Asian Conference on Computer Vision: ACCV*, 1998, pp. 672–679.

[5] V. A. Romer Rosales and S. Sclaroff, "3d hand pose reconstruction using specialized mappings," in *IEEE International Conference on Computer Vision: ICCV*, July 2001, pp. 378–385.

[6] D. Gavrila and L. Davis, "Tracking humans in action: A 3d model-based approach," in *IEEE Conference on Computer Vision and Pattern Recognition: CVPR*, 1996, pp. 73–80.

[7] Q. Delamarre and O. D. Faugeras, "3d articulated models and multi-view tracking with silhouettes," in *IEEE International Conference on Computer Vision: ICCV*, 1999, pp. 716–721. [Online]. Available: citeseer.nj.nec.com/article/delamarre99articulated.html

[8] T. K. G. Cheung, S. Baker, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *IEEE Conference on Computer Vision and Pattern Recognition: CVPR*, 2003.

[9] S. Ilic and P. Fua, "Generic deformable implicit mesh models for automated reconstruction," in *ICCV workshop on Higher-Level Knowledge in 3D Modelling and Motion Analysis*, 2003, pp. 29–38. [Online]. Available: citeseer.ist.psu.edu/ilic03generic.html

[10] M. B. R. Kehl and L. V. Gool, "Full body tracking from multiple views using stochastic sampling," in *IEEE Conference on Computer Vision and Pattern Recognition: CVPR*, June 2005, pp. 129–136.

[11] K. Komatsu, "Human skin model capable of natural shape variation," *The Visual Computer*, vol. 4, no. 3, pp. 265–271, 1988.

[12] P. J. Besl and N. D. Mckay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[13] K. Ogawara, K. Hashimoto, J. Takamatsu, and K. Ikeuchi, "Grasp recognition using a 3d articulated model and infrared images," in *Int. Conference on Inttelligent Robot and Systems*, 2003, pp. 1590–1595.

[14] M. D. Wheeler and K. Ikeuchi, "Sensor modeling, probabilistic hypothesis generation, and robust localization for object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 3, pp. 252–265, 1995.

[15] D. H. Thanarat Horprasert and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *ICCV Framerate Workshop*, 1999, pp. 1–19.

[16] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *PAMI*, vol. 16, no. 2, pp. 150–162, February 1994.

[17] C. Montani, R. Scateni, and R. Scopigno, "Discretized marching cubes," in *Visualization '94 Proceedings*, IEEE Computer Society. IEEE Computer Society Press, 1994, pp. 281–287. [Online]. Available: citeseer.nj.nec.com/montani94discretized.html