Generating 3-D Models of Human Motions by Motion Capture

I. Yamaguchi*, K. Tou*, J. K. Tan*, S. Ishikawa*, T. Naito**, M. Yokota**

* Department of Mechanical and Control Engineering, Kyushu Institute of Technology,
Sensuicho 1-1,Tobata, Kitakyushu 804-8550, JAPAN
(E-mail: {yamaguti, tou, etheltan, ishikawa}@is.cntl.kyutech.ac.jp)
** Department of Periodontology and Endodontology, Kyushu Dental College,
Manazuru 2-6-1,Kokurakita, Kitakyushu 803-8580, JAPAN

Abstract: A technique is presented for generating a compound human motion from its primitive motions obtained by a motion capture system. Some human fundamental motions are modeled in a 3-D way and registered as primitive motions. Because the factorization method is used for the motion capture, calibration of video cameras and connection of the motion in the direction of time is both unnecessary. Employing these motions, various compound human motions are generated by connecting the motions after having applied rotation and parallel transformation to them. Linear interpolation is done at the discontinuous boundary between primitive motions and smooth connection is achieved. Experimental results show satisfactory performance of the proposed technique. The technique may contribute to producing various complicated human motions without much effort using a strict motion capture system.

Keywords: Motion capture, 3D recovery, Compound motion, Modeling, Mixed reality

1. Introduction

Motion capture is a technique for producing three-dimensional (denoted as 3-D hereafter) models of human motions from the data that 3-D sensors such as a magnetic sensor and an image sensor provide. Recently, motion capturing techniques have been of much use in many fields including human motion analysis, 3-D human characters creation for amusement purposes, *etc*. This trend will be more and more enhanced because of increasing interests of our society in various human activities.

It is natural to connect some defined component motions in order to represent arbitrary human motions. To realize it, each component motion should be somehow adjusted so that it can be connected smoothly to a successive component motion. This is not very simple, however, when motion capture data has the description form of a set of 3-D coordinates of the feature points specified on a human body, since the detected number of feature points and the scale of the coordinates often depend on performed experiments.

This paper proposes a method of seamless connection of successive component motions described by a set of 3-D coordinates of specified feature points obtained from motion capture. The method realizes seamless connection first by superposing identical positions between the motions to be connected by the employment of a rotation matrix and a translation matrix, and then by performing interpolation between the successive frames composing the connected part of the motions. In this way, an arbitrary human motion can be modeled through combining by the proposed method 3-D component motions prepared in advance.

2. Developed Motion Capture Technique

In order to prepare for component motions, a developed motion capture technique[1] is employed in this study. This technique has some advantages over others in that calibration is not necessary with video taking cameras and that a stream of a component motion recovers simultaneously without 3-D data arrangement along the time axis. Moreover, the technique can be applied in any circumstances only if feature points on the person interested are detected on his/her video images.

Figure 1 shows an image capturing strategy in the present technique. F video cameras with fixed orientations are placed in front of the object interested. Neither their locations nor the orientations are employed in the technique though. Video camera f produces image stream $I_f(t)$ $(t=1,2,.....,T)$.

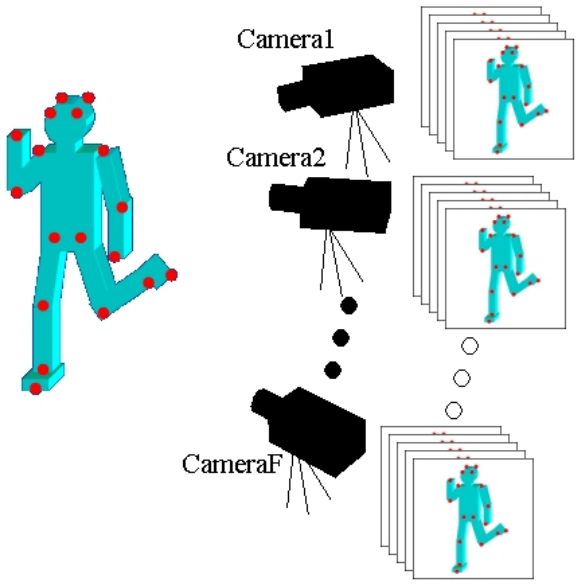


Fig.1 Configuration of a measurement system.

A feature point on the object at time t is denoted by $\boldsymbol{s}_p(t)$. Assume that the feature point $\boldsymbol{s}_p(t)$ be observed at the location $(x_{fp}(t), y_{fp}(t))$ on the frame of video camera f at time t. The set $R(t)$ is then defined by

$$R(t)=\{(x_{fp}(t), y_{fp}(t))|\ f=1,2,...,F; p=1,2,...P_t\}(t=1,2,...T).$$

It should be noted that the number of feature points P_t is allowed to be time dependent. This signifies that feature points may appear or disappear on an object and the object itself may emerge or vanish while taking its video image.

A $2F\times P_t$ matrix $W(t)$ is defined from the set $R(t)$ in the following form;

$$W(t)=\begin{bmatrix} x_{11}(t) & x_{12}(t) & ... & x_{1P_t}(t) \\ x_{21}(t) & x_{22}(t) & ... & x_{2P_t}(t) \\ ... & .... & ... & ... \\ x_{F1}(t) & x_{F2}(t) & ... & x_{FP_t}(t) \\ \hline y_{11}(t) & y_{12}(t) & ... & y_{1P_t}(t) \\ y_{21}(t) & y_{22}(t) & ... & y_{2P_t}(t) \\ ... & ... & ... & ... \\ y_{F1}(t) & y_{F2}(t) & ... & y_{FP_t}(t) \end{bmatrix} \quad (1)$$

The matrices $W(t)$ $(t=1,2,.....,T)$ defines a single matrix W of the size $2F\times Q$ as

$$W=(W(1)\,|\,W(2)\,|W(3)\,|\cdots|\,W(T)) \quad (2)$$

Here Q is the number of the entire feature points during the observation time, *i.e.*,

$$Q=\sum_{t=1}^{T}P_t \quad (3)$$

The extended measurement matrix of the size $2F\times Q$ is defined by

$$\tilde{W}=W-\frac{1}{Q}W\cdot E \quad (4)$$

where E is the $Q\times Q$ matrix whose entries are all unity. If the entry of matrix $\tilde{W}$ is denoted by $(\tilde{x}_{fp}(t), \tilde{y}_{fp}(t))$, we have

$$\tilde{x}_{fp}(t)=x_{fp}(t)-\frac{1}{Q}\sum_{t=1}^{T}\sum_{p=1}^{P_t}x_{fp}(t)$$
$$\tilde{y}_{fp}(t)=y_{fp}(t)-\frac{1}{Q}\sum_{t=1}^{T}\sum_{p=1}^{P_t}y_{fp}(t) \quad (5)$$

Let us project all the P_t feature points at time t $(t=1,2,.....,T)$ onto the 3-D space at $t=1$ and take the origin O of the 3-D space at the center of all the Q feature points there. If orthographic projection is assumed, we have [3]

$$\tilde{x}_{fp}(t)=(\boldsymbol{i}_f, \boldsymbol{s}_p(t))$$
$$\tilde{y}_{fp}(t)=(\boldsymbol{j}_f, \boldsymbol{s}_p(t)) \quad (6)$$

where $\boldsymbol{i}_f$ and $\boldsymbol{j}_f$ are the unit column vectors in the horizontal and vertical orientation, respectively, defining the coordinate system on the image plane of video camera f and the unit column vector $\boldsymbol{k}_f$ defined by $\boldsymbol{k}_f=\boldsymbol{i}_f\times\boldsymbol{j}_f$ coincides with the light axis of the video camera. Projected location of the origin O onto the image plane of video camera f is given by the second terms of the right-hand side of Eq.(5).

From Eqs.(4) and (6), the matrix $\tilde{W}$ is separated into two matrices in the form

$$\tilde{W}=MS \quad (7)$$

where M is a $2F\times 3$ matrix of the form

$$M=\begin{pmatrix} \boldsymbol{i}_1^T \\ \boldsymbol{i}_2^T \\ \mathbf{M} \\ \boldsymbol{i}_F^T \\ \boldsymbol{j}_1^T \\ \boldsymbol{j}_2^T \\ \mathbf{M} \\ \boldsymbol{j}_F^T \end{pmatrix} \quad (8)$$

giving video camera orientations and S is a $3\times Q$ matrix of the form

$$S=(\boldsymbol{s}_1(1), \boldsymbol{s}_2(1), \boldsymbol{s}_3(1), \mathbf{K}, \boldsymbol{s}_{P_1}(1)\,|\,\boldsymbol{s}_1(2), \boldsymbol{s}_2(2), \boldsymbol{s}_3(2), \mathbf{K}, \boldsymbol{s}_{P_2}(2)\,|\,\mathbf{L}\,|\,\boldsymbol{s}_1(T), \boldsymbol{s}_2(T), \boldsymbol{s}_3(T), \mathbf{K}, \boldsymbol{s}_{P_T}(T)) \quad (9)$$

Shape matrix S contains all recovered feature points. The set of the recovered feature points at time t $(t=1,2,.....,T)$ is denoted by $S(t)$, *i.e.*,

$$S(t)=\{\boldsymbol{s}_p(t)|\ p=1,2,\cdots,P_t\}. \quad (10)$$

3. Obtaining 3-D Recovery Data

Motions of a person were taken images by 3 fixed video cameras. Three views of a subject taken by the three video cameras are shown in **Fig.2**. The number of feature points adhered on the person was 17 or 18. The obtained video images were stored into a memory as a set of still images containing 24 frames per second. From these stored images, the extended measurement matrix $\tilde{W}$ of Eq.(5) was defined and, by applying Singular Value Decomposition (SVD) to it, the shape matrix S was obtained. Computation time for recovering a 10 second motion (240 frames) with 18 feature points was approximately 5 seconds by a PC equipped with a PentiumⅢ processor (1.2GHz) as a CPU. Part of the result is given in **Fig.3**, where human shape is represented by a set of planes instead of a set of feature points for understandability. As this result was made using OpenGL, it is possible to observe it from any angle by rotation.

4. Connecting Motions

The employed motion capture technique doesn't necessitate camera calibration. Instead one cannot specify the 3-D coordinate system in which locations of recovered feature points are described. This means that two motions recovered by individual SVD application are described in the respective 3-D coordinate systems with possible scale difference. Therefore transformation between the two coordinate systems is indispensable to realize seamless connection of successive motions. By the seamless alignment of individual motions, a model of any complicated motion can be produced.

Employed format of the 3-D recovered data is the 3-D coordinates. By these data, a transformation matrix is defined containing translation, rotation, and scaling in order to superpose one coordinate system to the other. The rotating matrix, the parallel transformation matrix, and the scaling matrix are shown below. If a feature point is denoted by $\mathbf{C}(x, y, z, 1)$, it is converted to $\mathbf{C}'$ by multiplying the transformation matrix $\mathbf{T}$. The transformation matrix is a 4 by 4 matrix, because homogeneous coordinates are used.

$$\mathbf{C} = \begin{pmatrix} x & y & z & 1 \end{pmatrix} \qquad \mathbf{C}' = \mathbf{C}\mathbf{T} \tag{11}$$

Rotation around X axis:
$$\mathbf{T}_x = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos q & \sin q & 0 \\ 0 & -\sin q & \cos q & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{12}$$

Rotation around y axis:
$$\mathbf{T}_y = \begin{pmatrix} \cos q & 0 & -\sin q & 0 \\ 0 & 1 & 0 & 0 \\ \sin q & 0 & \cos q & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{13}$$

Rotation around Z axis:
$$\mathbf{T}_z = \begin{pmatrix} \cos q & \sin q & 0 & 0 \\ -\sin q & \cos q & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{14}$$

Parallel transformaion:
$$\mathbf{T}_p = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ P_x & P_y & P_z & 1 \end{pmatrix} \tag{15}$$

Scaling transformation:
$$\mathbf{T}_s = \begin{pmatrix} S_x & 0 & 0 & 0 \\ 0 & S_y & 0 & 0 \\ 0 & 0 & S_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{16}$$

If the number of recovered feature points is mutually different between two recovered motions, the coordinates of the missing feature points are interpolated in a 3-D way to make the number identical. To realize further smoothness, two successive motions are interpolated between the end of the first motion and the beginning of the successive motion. Linear interpolation is employed for this purpose.

In the performed experiment, a standing-up motion from the chair is followed by a raising-hand motion. Initially the respective motions recover three-dimensionally in separate experiments using the developed motion capture technique. They are then applied the seamless connection in the explained way. The result was satisfactory. In the combined motion, a human model stood up from a chair and raised his right hand without loosing smoothness of the entire motion.

5. Discussion and Conclusions

The paper proposed a method of realizing seamless connection of two separately recovered motions. Various compound human motions were generated by connecting the motions after having applied scaling, rotation, and parallel transformation to them. Linear interpolation was done in the discontinuous boundary between primitive motions and smooth connection was achieved for new motions. Performance of the method was certified experimentally. The method may contribute to the production of 3-D models of various motions from prepared component or basic motions. This signifies that the method imposes a user less laborious work of motion capture, since he may make use of other 3-D motion models already recovered till then. The developed method will be employed for creating a 3-D virtual human as a man-machine interface [2].

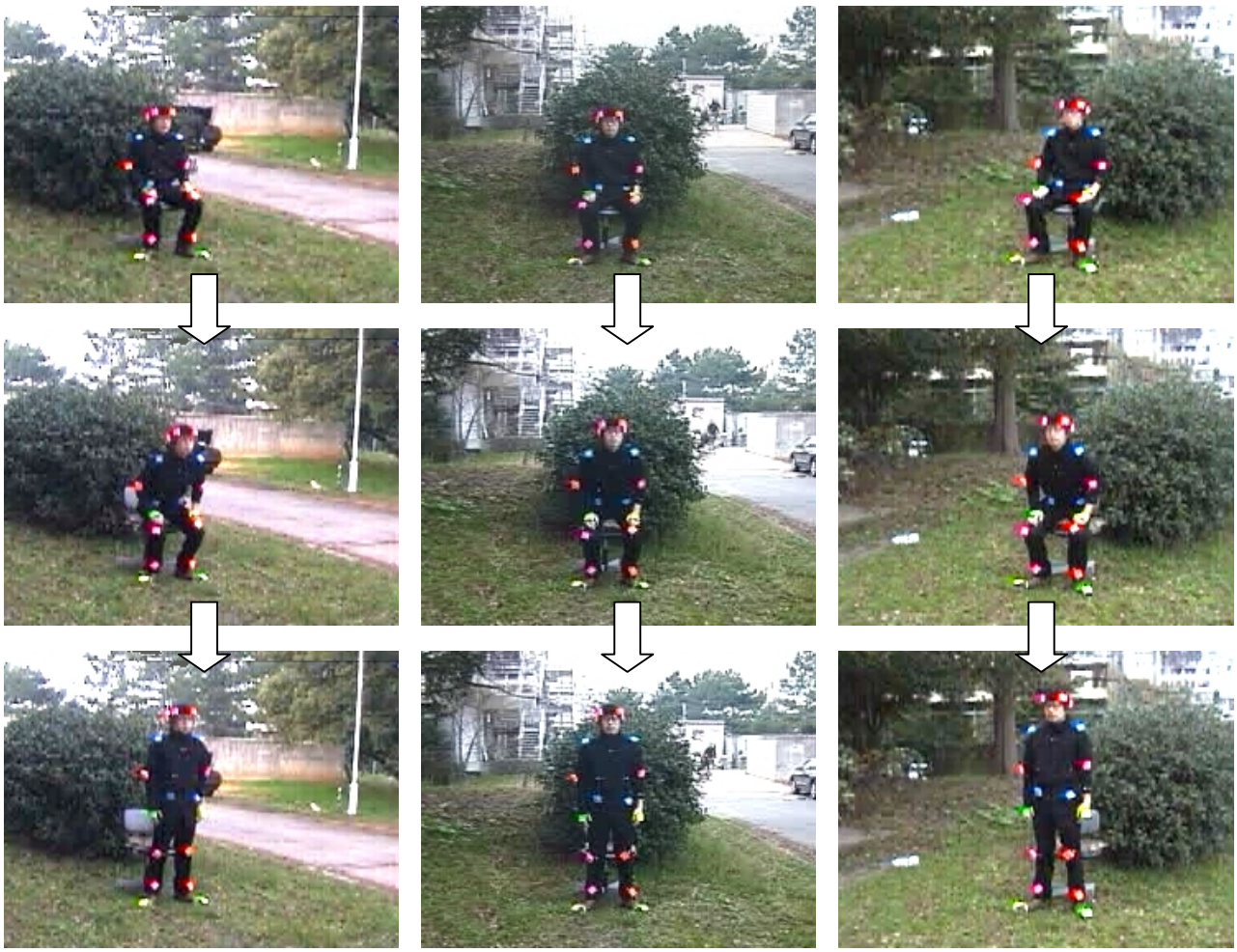
A 3D virtual human made by this technique is planned fused into the real space by using the method of Mixed Reality. An interactive system intended for improving human work efficiency will be constructed in future by displaying a virtual human made by the proposed technique in the real space in front of a user and giving useful information to him/her. Examples may include the operation explanation of various instruments by 3-D CG character, and a guide system by a virtual character taking a guest to his/her destination using GPS, *etc.*

References

[1] J. K. Tan, S. Ishikawa : "Human motion recovery by the factorization based on a spatio-temporal measurement matrix", *Computer Vision and Image Understanding*, Vol.82, No2, pp.101-109 (2001).

[2] D. Hirohashi, , J. K. Tan, , S. Ishikawa,: "Producing a virtual object with realistic motion for a mixed reality space", *Proc. International Conference on Control, Automation, and Systems*, pp.1084-1087 (2001).

[3] C. Tomasi, T. Kanade: "Shape and motion from image streams under orthography: A factorization method, *Int. J. of Computer Vision*, Vol.9, No.2, pp.137-154 (1992).



(a)The left views (b)The center views (c)The right views

Fig.2 Views of a subject taken by three cameras

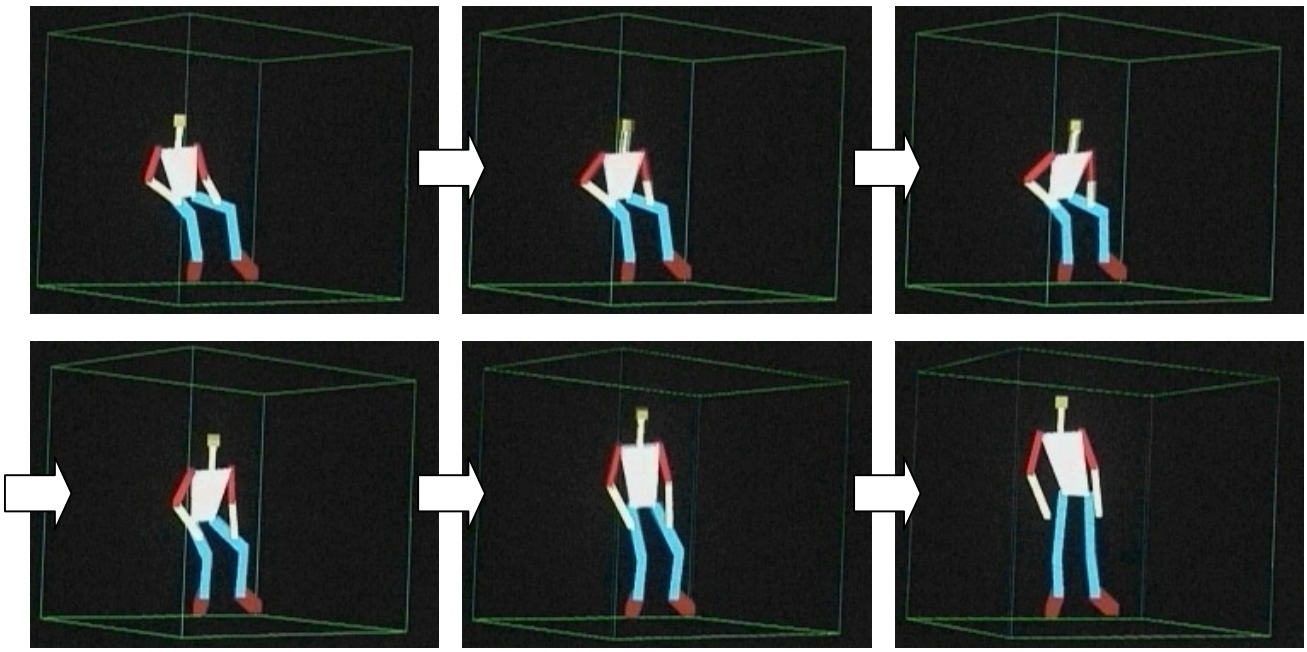


Fig.3 Recovered motion samples