Interactive Human Motion Acquisition from Video Sequences

Jiang Yu ZHENG, Shigeru SUEZAKI, Yasuhiro SHIOTA
Kyushu Institute of Technology
Iizuka, Fukuoka 802-8500, Japan
zheng@mse.kyutech.ac.jp

Human motion modeling for animation and VR has reached a level of capturing real motion data from people using various visual and non-visual sensors. However, most of the available systems require special devices and environment, or even attach extra things onto an actor. To realize a personal system, we developed a general type software to acquire arbitrary motion from a video sequence. A 3D articulate human model with changeable size, color and surface shape is constructed and personalized to fit with a focused figure in the video. The personal model is then driven either automatically or manually to match with the moving body in the consecutive image frames. This matching starts from key frames that contain key poses. A smooth motion is then interpolated in between key frames. And the evaluation of the generated motion is enhanced by image correlation. We provide various methods to make the matching feasible in order to reduce the modeling time. This approach is suitable for personal use to meet wide needs of human motion acquisition.

1. Introduction

Extracting human motion has become a hot topic in recent years as virtual reality, multimedia, and computer interface technologies are developed rapidly. Generating a motion sequence is no longer tuning some predefined oscillating function mathematically, but to capture real actions of humans which not only yields natural movements but also increases motion variations. The obtained motion sequences can be applied to a sophisticated human CG model to realize virtual acting in entertainment, and be displayed in different viewing directions for education and training purposes. However, many current systems attach sensors onto a player physically or require wearing a kind of body suits, which limits the movable space. Besides the device-based approach using position sensors and suits, there is also a visual sensing approach that measures human motion at each time instance. Mark-tracking approach employs multiple cameras that locate LED or balls on articulators of the human during the performance. Other approaches extract 3D shape of human at each time instance [1,2,3]. Although this type of systems has more free space for a player to move, it still needs a studio environment with controllable illumination, background, and camera positions. Moreover, they are usually equipped with powerful devices for processing multiple image sequences from a number of cameras. Complicated camera calibrations are also required. None of the above method is capable of catching motion parameters from a past performance recorded in old movies or videos.

What this work aimed is a personal system. It can catch human movement from a video sequence. If various kind of motion could be edited and generated at a personal level, individuals would be able to upload their resulted sequences of motion parameters to a database via internet so that a big motion base will appear on the net for various uses.

As a merit of our approach, we use images from one camera. It is the most common and flexible way to record human activity. The video sequences including pan and zoom can be obtained from TV shows, home videos, old films, etc. Arbitrary motions at any place are allowed to appear in the sequences.

Along this line, many works in computer vision have been done to recognize gestures. They do image segmentation guided by body constraints for certain kinds of specific motions. Most of them were implemented and tuned on selected lab sequences, and have no insurance in dealing with arbitrary, outdoor human motion. The reason is mainly the existing gap between pixel patterns and semantic descriptions of human motion in the processing process [7]. Adding stereo and more cameras increases much computational cost but not much information.

It has been known that model-based approach can save many computations and devices [4, 6, 8]. We choose the model-based approach to extract motion from an image sequence taken by a camera. Rather than relying on an automatic extracting system, we developed a software-based system that can interactively generate various kinds of motions. A 3D articulate human model with

0-7695-0643-7/00 $10.00 © 2000 IEEE

changeable sizes, proportions, colors and surface shapes is prepared to facilitate matching between the model and image figures. By referencing one or several particular images in the sequence, we adjust the model through an interface to produce a personal model of an image figure. The motion generation of the figure starts from obtaining a set of key frames [9,12]. The personal model is matched with the images to determine key poses; each part of the body is aligned with the figure in the background image. The motions between key frames are predicted and interpolated. The actual traces are verified by color based correlation between the model and images. We test various sequences and it achieved a stable extraction of human motion.

The first part of this work is to build an adjustable 3D model of human called standard model to facilitate model-based approach. It contains a function of articulate motion and functions of changing size, proportion, shape and color of the body. It can be modified to fit different figures in the images. After a sequence of images has been taken, we select one or several frames in which the focused figure has a state of stretching body. By moving the body of the standard model and adjusting its articulation to overlap with the image figure, the personal model sizes are produced. Also, with this fitting, we determine the color on the model from the image so that a color-based correlation can be applied later in the motion verification.

The second part is the matching between the adjusted personal model and the figure in the whole video sequence in order to extract a flow of articulation parameters. The matching begins with manual determination of key frames where the human motion has an obvious change in directions. We move the personal model and align it over the figure in each key frames. It is followed with an automatic interpolation and verification of poses in the remaining frames. The prediction and interpolation of intervening motion is made based on motion smoothness. The pose of the graphics is substantially generated model at each instance for correlation with the image, which is more robust than frame to frame matching. This top-down approach provides a visible and clear guidance to the matching process. With many convenient interactions between an operator and the model, our system can almost cope with any kind of human motion.

The interested motions range from dance and performance on the stage to various indoor and outdoor sports. Images are either from TV broadcasting or a personal video camera, which may have pan and zoom changes. We have no particular demands on background and environment illumination. Our objective is to obtain distinct and smooth motion from an image sequence for animation and other multimedia application. Rather than obtaining accurate parameters, we pay more attention on the visual appearance of the extracted motion sequences. Detail motions occluded by clothes and motion hard to be observed in the images are ignored.

2. Standard Model of Humans

What kind of model is necessary for guiding motion extraction? Some articulate function models have been reported [4, 6], in which optical flow or moving edges are constrained by the models. These simple skeleton models and edge-based guiding approach, however, lacks of information on shape and color of a human. It may loss tracking if edges are failed in extraction due to the influence of background and occlusion. The flow-based approach is actually a matching between consecutive frames guided by the model. This method is not only week in noise because the differentiation between video frames is computed, but also complex in computation.

This work generates a volumetric model with color to provide more information for tracking human motion [5]. A graphics model avoids a complex computation of the relation between joint angles and body positions. It is easy to handle in a real application. Moreover, the matching is between the model and figures, which is more robust than those using differential information between frames. With the progress of CG and GUI, building a model or an auxiliary structure becomes more convenient than before. The difference of this work from previous ones lies in that our model not only serves as a constraint, but also works as a reference of matching; comparing color of the model with the image becomes possible. Rather than employing more devices or adding algorithms, this work puts more effort on improving the model and uses more information from the model in motion tracking.

First, we build a standard human model according to the anatomy proportions (Fig. 1). The model has three functions.

1). The model is articulately moveable.

2). The size of each part of the model is changeable.

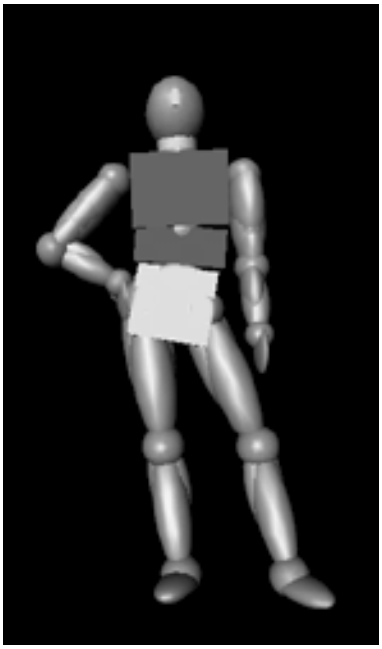
3). The model can be colored or textured.

4). The whole model can be scaled, translated and rotated.

The center of the model, or the coordinates system of the model, is located at top of the chest and center of two shoulders. The 3D position of the model is then registered as (X(t), Y(t), Z(t), Rx(t), Ry(t), Rz(t)) in the world (window) based coordinate system, where t is the time.

The model is in a tree structure described by using Open Inventor software. The structure has 16 articulators at head, neck, back, waist, shoulders, elbows, wrists, hips, knees, ankles, which are denoted by Ai(t), i=1,…,16. Each articulator can be rotated either automatically or manually in three degrees of freedom; rotates around x, y, z axes respectively in their local coordinate systems. Each

rotation is constrained in certain range based on the body structure. The three degrees of rotation can also be represented as an angle ϕi around a changeable axis $\mathbf{V}i$. In our graphics interface, auxiliary spheres are located at articulators: clicking any of them turns on a control ball for manual adjustment of articulate rotation (Fig. 2). This allows us to manually create any kind of pose of human body.



(a)

(b)

Fig. 1 Model structure containing articulators and inflexible parts. Identical parts are put on both left and right limbs. (a) Geometric parts are combined into inflexible parts. (b) Model covered with segmented smooth surfaces.

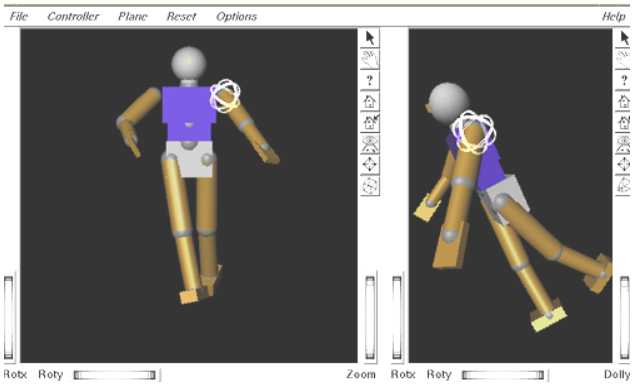


Fig. 2 A graphics interface for manual adjustment of the standard model. Each articulator can be selected and then rotated by controlling a sphere attached to it after being selected. Two windows are displayed in the interface for an easy manipulation of the same model.

Between articulators are inflexible parts including the head, the neck, arms, forearms, hands, the chest, the stomach, the hip, thighs, calves, and feet. Their sizes can be changed relative to the head height in three directions. The whole model can also be enlarged and squeezed. Changing the head height scales the entire model size. The sizes and proportions can be controlled in an interactive way through GUI so that the model can be fitted to various figures in the images (Fig. 3). For every inflexible part, it has three scale factors along its own x, y, z directions. Scaling these parameters using the slide bars on the interface window can change the length, width, and thickness of that part.

There are two ways to construct inflexible parts in our system. The first choice is to use multiple ellipses or cubes that represent muscles. Because of the simple data structure, the composed model runs fast in display. The drawback is that the color can only be defined to the level of each part. A centralized RGB value sampled from the image is assigned to each part, which is roughly enough in coarse matching when the figure is small in the images. Painting a small image and mapping it onto a part can generate body texture such as different costume marks. Another way is to build a sophisticated surface model. We generate a surface model by using commercial products and truncate them into inflexible parts as shown in Fig.1.b. Each part then contains many triangular patches. We can even substitute an inflexible part with surfaces obtained densely from a laser range finder, if the graphics machine is powerful enough to display a large amount of patches. In these cases, the model color can be defined to a detailed level of patches.

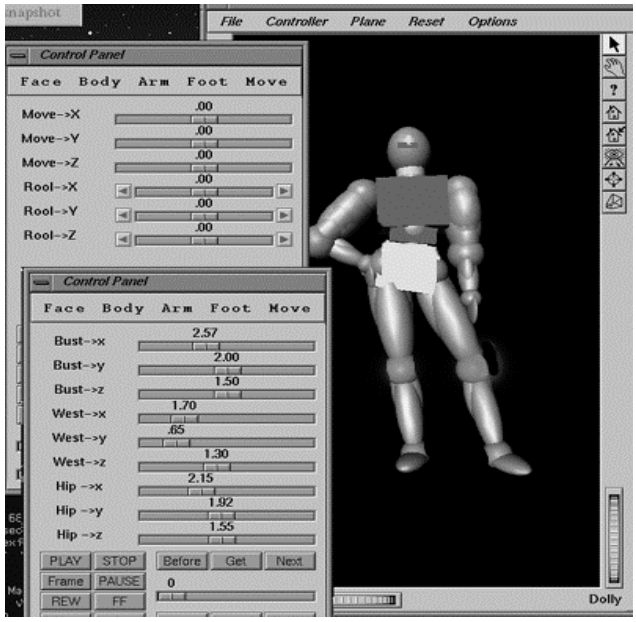


Fig. 3 Inflexible parts can be changed in their sizes, shapes and colors using a graphics interface.

3. Generating Personal Models

Next phase is to prepare a personal model from the standard model using one selected image from the video; the focused human is better to extend his/her limbs and trunk in the image. This is also done using the graphics interface, in which the image is displayed on a background panel behind the model under perspective projection (Fig. 4). It is a texture mapping of the image onto a plane parallel to the computer screen. Selecting a proper Z value, the standard model is moved and fitted to an acting person in the image by repeatedly using following steps:

1. Determining $X(\tau)$ and $Y(\tau)$ positions, and rotations $Rx(\tau)$, $Ry(\tau)$, $Rz(\tau)$ of the model by referring the chest.
2. Changing the entire size of the model by scaling the head.
3. Rotating each articulation of the body to produce the pose of the moving person in the image,
4. Scaling the length, width, and thickness of each inflexible part to fit with the figure in the image.

In order to implement above steps efficiently, two windows in the interface provide different views of the model for comparing and generating the pose in the image. Clicking a control ball at an articulator and rotating it can pose parts under that articulator. The operation can also be performed in the second window if the control ball in the first window is hard to click and drag from the particular viewing direction. The model can even be tuned half-transparent to give us a clear view on both the model and the underlying figure. The obtained size parameters of the model are registered as the personal model, which mainly contains information on inflexible parts. Figure 5 shows an example of the personal model.

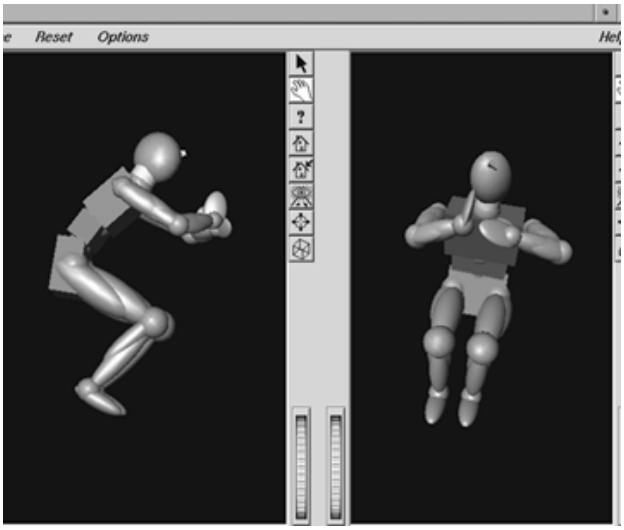


Fig. 5 One example of personal model obtained from deforming the standard model, using an image in a diving sequence.

On the bases of shape and size, we further need to catch texture on the body. If the costume is simple or the image size of the human is small, we only need to get a centralized parameter from each image region of the body and assign it to the corresponding inflexible part on the model. If there are complicated patterns on the body, the average is computed in that region for the color value. The balls located at articulators for popping up control spheres can be set to the same colors as the closest inflexible parts.

If the figure in the image is large, a surface model is preferred to use. We project each patch onto the image plane and average the image color in the projected region. It is then assigned to the patch as its texture value. A surface model is usually heavy in running compared with the cubes and ellipses composed model. But it contains more details of the body texture and will be useful in matching a model with large figures in the images.

4. Determining Key Poses

For many sports and performance scenes, a complete automatic approach is hard to succeed on the whole sequence. We select key frames from the sequence to manually locate the model. These key frames are chosen in the following way

(a) A key frame is set when a drastic move starts or a stroke of action finishes for major parts of the body; the

rotations of many articulators are not smooth at such a moment.
(b) A frame from which the camera starts zooming or pan is selected as a key frame,
(c) If a body rotation is over 180 degree, a key frame is added to indicate the rotating direction.

How many key frames selected depends on the complexity of the action of interest. Since the manual fitting of the model to an image is very easy using the developed interface, the number of key frames is not a crucial problem.

With the personal model and the whole image sequence, we fit the model to the human occupied region in the images for motion estimation. A video sequence for several seconds having about 200 images is usually taken. We use a device that can control the video and take images frame by frame into the computer. Images can also be saved in hard disc space if necessary.

When an image sequence has been taken, the camera might have some pan and zoom. The image position and distance of the figure may also change during the movement. Among them, a zoom or a big variation in distance will make the human size change in the images. These two effects are compensated in the key frame alignment by changing one parameter S(t) which is the model size. This is because not in all the image sequences the values of zoom and distance can be easily figured out simultaneously. The output of pose parameters at each instance includes model position (X(t), Y(t), Z), orientation (Rx(t), Ry(t), Rz(t)), model size S(t), and rotation parameters ($\phi i(t)$, $\mathbf{V}i(t)$) of 16 articulators. The motion thus is described by a flow of pose parameters. When an obtained motion sequence is replayed with a different model, we can even neglect its translation (X(t), Y(t)) and model size S(t) so that the model is moving at the same position without shifting in the frame.

The interface window can also show information of the current frame, move frame forward and backward, and jump to a particular frame randomly (Fig.3). The generated motion parameters can be saved to a file and be loaded again to the interface to replay the motion. These operations can be done either for a whole sequence or for a single frame.

3

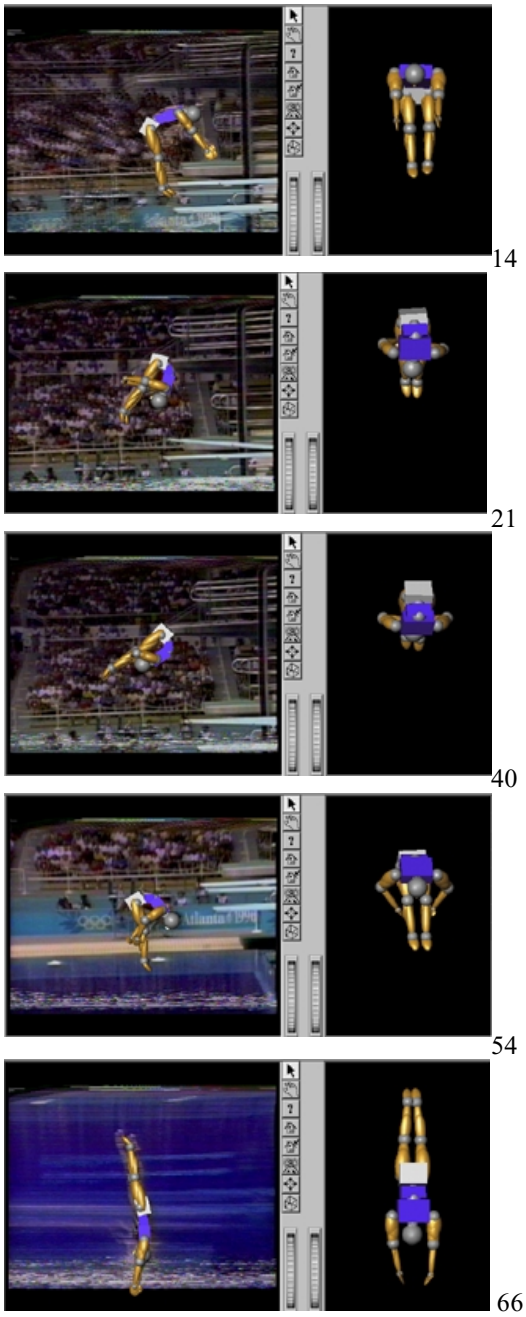


Fig. 6 Matching the personal model with key images in the background. The opacity of the model can be changed to show the image behind.

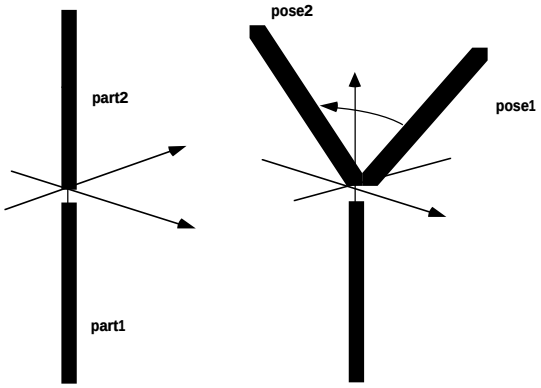
The matching between the personal model and the key frames is in the same way as adjusting the standard model. The exception is at the background image scaling if the camera had a zoom in taking images. Since the lengths of limbs and trunk of a personal model have been fixed in the previous steps, which provides a strong constraint, the 3D position of an inflexible part can be

determined by achieving a best overlapping. Because one image has determined 2D direction of a limb, the known length of the limb only yields two possible positions in depth. We can simply choose a proper one from them. The alignment of the whole body is carried out along the tree structure of the body, which starts from the chest, and then sequentially goes to neck and head, goes to hand through arm, forearm, and goes to leg and feet through trunk, respectively

Figure 6 shows a series of key frames matched with the model. There is some occluded part invisible during the operation. We determine its motion parameters based on the obvious pose in the images.

5. Automatic Generation of Other Frames

Now, we need to determine the consecutive poses of the model between key frames. There are many choices in determining rotation values of all the articulators. Sixteen articulators compose a parameter space of 48 degrees that is impossible for searching optimal values. Even the rotation limits of each articulator are used as constraints, it only confine the search scope to half or one third. To solve this problem, the rotation angle of each articulator is interpolated to produce an approximate motion trace. This trace should be smooth according to the definition of a key frame. Assume the poses of articulator *i* at two consecutive key frames t and t' are $\phi(i,t)$ and $\phi(i,t')$ around axes **V**i(t) and **V**i(t') respectively, a spherical linear interpolation will make an output smooth for both the rotation angle and axis (Fig. 7). The generated parameters can even be a simple non-linear function of time according to psychology and physiology experiment on human motion. The motion trace of each part is usually non-linear simply because the motion involves articulate rotation. Model size S(t) and its position (X(t), Y(t)) are interpolated linearly between two key frames.



(a) an articulator. (b) pose t to pose t'

Fig. 7 Representation of an articulor and spherical linear interpolation between two poses.

The evaluation of the personal model with an image body is simply done by color correlation between the model-generated image and the real image overlapped. In generating a continuous pose sequence between two key frames, the 2D projection of the model is obtained at each instance to match with the corresponding image. We use the correlation as follows to move out the influence from general intensity changes in the video frames.

Suppose M(t) is a model occupied region with size Sm(t) which slightly changes in different frames, and p is a 2D image point in region M(t), the correlation C(t) is evaluated by

$$C(t) = \frac{\sum_{p \in M(t)} \left(img(t,p) - \mathrm{mod}(t,p) \right)^2}{S(t)_m^2}$$

where mod(t, p) is the color of point p projected from the 3D model.

If the correlation value monotonically increases from an average value, we conclude that the predicted motion diverts from the real motion in the images. This divergence is caused by insufficiency of key frames. The system reports the missing of key frame and a new key frame will be inserted for the motion trace interpolation. A key frame can be inserted anywhere in the sequence if necessary. The interpolations and verifications are applied again between the entered key frame and its neighboring key frames previously set.

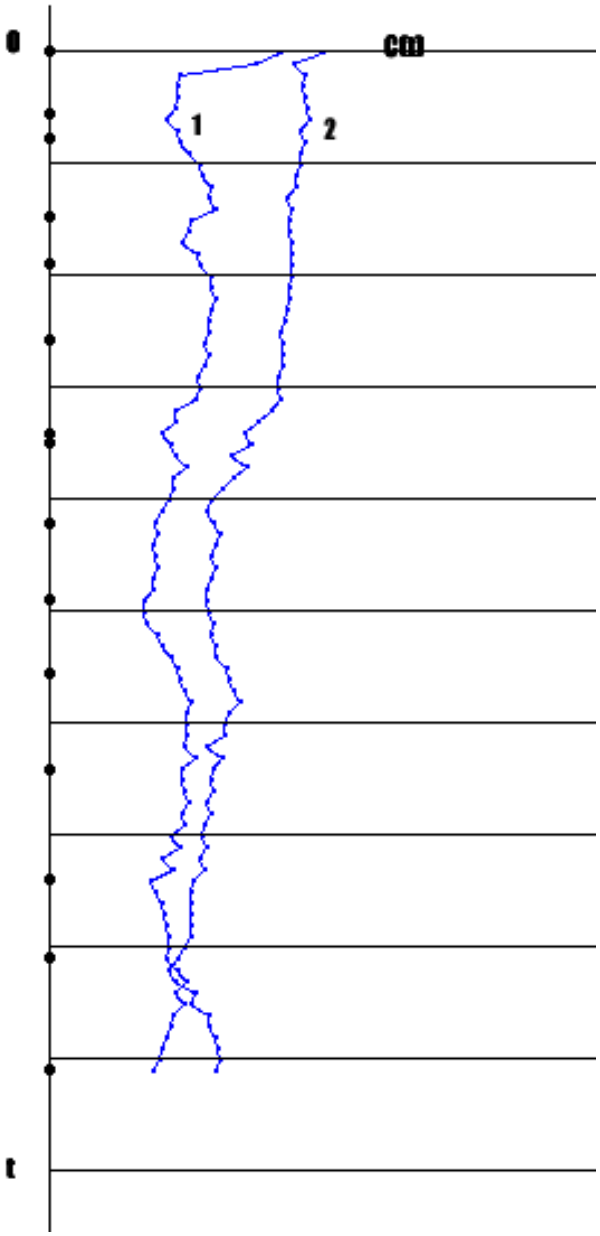
If the resulted distribution of correlation gets worse in the middle of two key frames, but keeps good near the key frames, we change interpolation slightly by shifting the trace along the time axis. Because a motion stroke is probably fast at the beginning and slows down close to the end, a simple non-linear shifting of the motion parameters along the time axis works well.

6. Experiment and Discussion

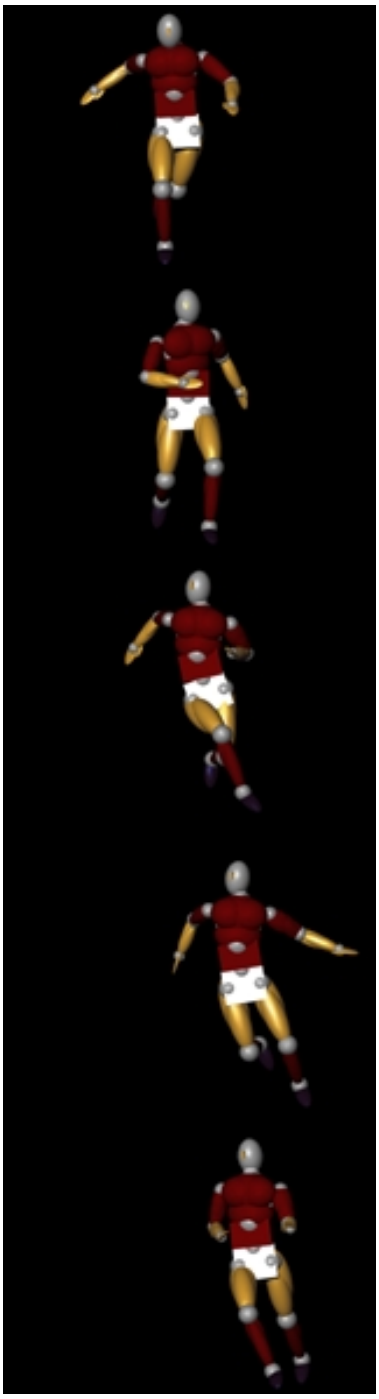
Images are taken into a hard disc frame by frame using a video controller (Sony Vbox). Only odd or even lines in the interlaced frame are selected for avoiding motion blur. SGI O2 machine is used for graphics and interface. We use Open Inventor graphics package to display 3D models and images, motif for designed interface, and x window for dynamic image display and correlation. Editing a complex motion of a few seconds (30 frame/sec) usually takes an hour. It is certainly depend on skill of the operator and the machine speed. Figure 8 gives the result of the diving sequence and it can also be found in the web site [11].

Figure 9 is another example of this approach working on ballet. Totally, 98 frames are taken and finally 20 key frames are selected (some of them are displayed in the top raw). The whole sequence of the motion is generated and some of them are displayed

below. Figure 10 displays a sequence of correlation value (zero at the top). New key frames are added into the sequence at where the correlation tends worse in a period (excluding a specific instance such as a camera flash is on from the background). Finally, we are able to keep the correlation value small in the whole period. We have tested many sequences (Fig 11) and found this approach is a relatively practical way in doing motion extraction for multimedia application. Some experiment sequences of human motion including short clips of football (700KB), dancing (1MB), and skating (300KB) scenes are provided in QuickTime format at web site for reference.



(a)



(b)

Fig. 10 A football sequence displayed with the value of correlation. (a) Correlation distributions of the whole sequence. Dots indicate the position of key frames. Curver1 is from model poses well fit to the images. Curve 2 starts from a non-matched position and finished at well fit positions. We can notice that curve 2 is worse than curve 1. For the best fitting poses, the correlation keeps a low value. (b) Samples of the generated sequence.

Fig. 11 One of the video clips in the following for motion generation: diving (96 Oylimpic champion diving), football shooting (99 world cup), dance (Russian ballet), and skateing (96 Winter Oylimpic man's champion). The video clips could be obtained at web site [11].

Is the described model sufficient enough? Although it is not as fine as those generated by high level graphics, it serves as a dynamic template very well for small figures in a rapid motion. The accuracy of the extracted motion is evaluated on the level whether the reproduced motion sequence is visually close to the original image sequence. The correlation does not work when a human body has complex texture similar with the background. It may also fail in an image sequence that has low saturation or with many areas of shadow.

The number of key frames is still large if we want to achieve a fine and realistic movement. It is our next objective to reduce key frames in manual adjustment so that the system can work more efficiently. Also, we find that determining a twist of body is not easy, because this kind of rotation of inflexible parts, particularly under clothes, does not provide much visual information. This problem also happens in other automatic vision approaches. The loss of twist value on limbs is not a big problem because the generated poses can still look normal. However, a miss-estimation in twist of trunk may inference the limb locations.

7. Conclusion

This paper introduced a tool to extract human motion from a video sequence for animation and other multimedia applications. A graphics human model is used in matching with the images. A sophisticated model will provide more opportunities to fit with images and reduce uncertainty in the human motion understanding. We first produce a 3D personal model according to one selected image in the video sequence. Then, we match the personal articulate model with the entire image sequence to obtain a vector of motion parameters including body positions and articulator rotations. Key frames are manually selected, and automatic interpolation and verification are performed. The motion parameters will be used to drive various personal models for generating dynamic and virtual scenes. It achieved a personal system so that people can obtain motion more freely from daily life.

8. References

[1] S. B. Kang, J. Zitnick, T. Kanade, A multibaseline stereo system with active illumination and real-time image acquisition, ICCV95, pp. 88-93, 1995.

[2] A. Katkere, S. Moezzi, D. Kuramura, P. Kelly, R. Jain, Toward video-based immersive environments, Multimedia Systems, Vol. 5, No. 2, pp. 69-85, 1997.

[3] S. Moezzi, L.Tai, P. Gerard, Virtual view generation for 3D digital video. IEEE Multimedia.Vol. 4, No. 1, pp.18-26, 1997.

[4] M. Yamamoto, K. Koshikawa, Human motion analysis based on a robot arm model, CVPR91, pp. 664-665, 1991.

[5] P. Beylot, P. Gingins, P. Kalra, M. Thalmann, N. Thalmann, J. Fasel, 3D interactive topological modeling using visible human dataset, Eurographics 96, Vol. 15, No. 3, pp. C-33C-44, 1996.

[6] K. Rohr, Towards model-based recognition of human movements in image sequence, CVGIP Image understanding, Vol. 59, no. 1, pp. 94-115, 1994.

[7] Q. Chen, H. Wu, M. Yachida, Face detection by fuzzy pattern matching, ICCV95, pp. 591-595, 1995.

[8] J. Y. Zheng, S, Suezaki, A Model Based Approach in Extracting and Generating Human Motion, 14th ICPR, Vol. 2, pp. 1201-1205, 1998.

[9] Y. Ohta, T. Yamagiwa, M. Yamamoto, Keyframe tracking of human body in 3D motion from a monocular image sequence. Trans. IEICE, D-II, Vol. J81-D-II, No. 9, pp. 2008-2018, 1998.

[10] I. A. Kakadiaris, D. Metaxas, Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection, CVPR96, pp.81-87, 1996

[11] http://sein.mse.kyutech.ac.jp/~zheng/human.html

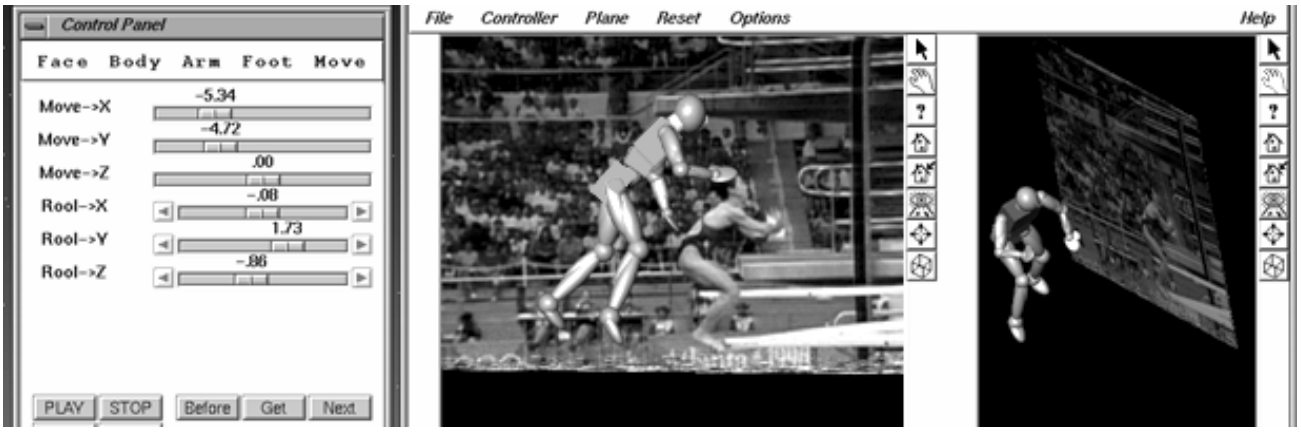


Fig. 4 Moving a standard model over a figure in the image and adjusting the model size to produce a personal model. Left: a control panel of the 3D model that can change parameters of the model. Middle: an image of an Olympic diver behind the model for model adjustment. Right: another view of the model and the image plane for a free manipulation of the model.

Fig. 8 Extracted motion sequence of the diver. Totally about 120 images are generated and the results are displayed at every 6 frames.

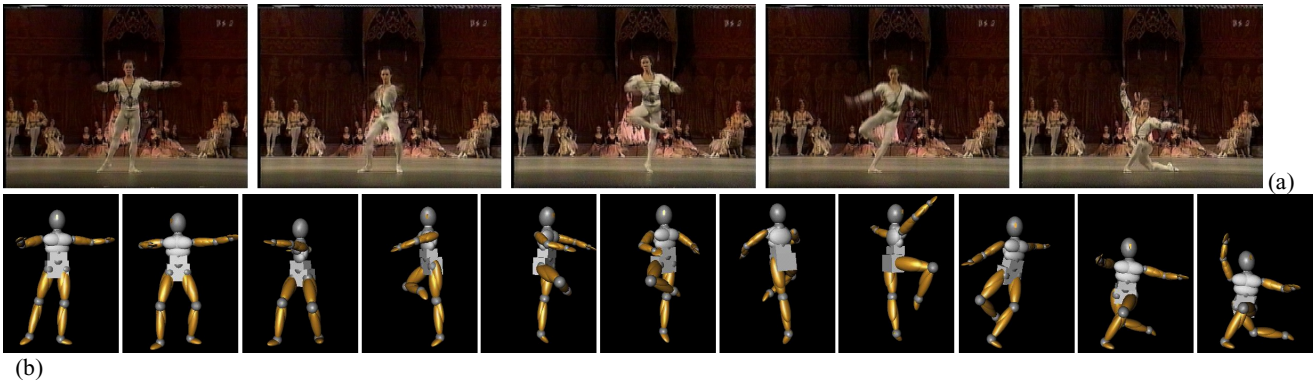


Fig. 9 An example of generating the whole sequence of motion from key frames. (a): several selected key frames in the image sequence. (b): frames of generated motion displayed with the model.