Human Skeleton Modeling from 2D Uncalibrated Monocular Data

En Peng
Department of Computing
Curtin University of Technology
GPO Box U1987 Perth
6845, Perth, Australia
perryp@cs.curtin.edu.au

Ling Li
Department of Computing
Curtin University of Technology
GPO Box U1987 Perth
6845, Perth, Australia
ling@cs.curtin.edu.au

ABSTRACT

A novel method which is simple but effective is proposed to estimate human skeleton ratios from 2D uncalibrated monocular data. Unlike the existing skeleton estimation methods where pre-defined models are used or identification of body segments with certain attributes is necessary, the proposed method utilizes only the 2D joint locations as the input source without posture estimation. In addition, the proposed method uses a real perspective camera model instead of the popularly-used scaled orthographic camera model. The proposed method is tested on monocular data from different camera motions and the estimation result is satisfactory. The reconstructed human skeleton model can be used in further research for full body reconstruction or motion reconstruction from monocular data.

Keywords

Modeling, human skeleton proportion, monocular data

1. INTRODUCTION

Applications involving virtual human, such as digital movies, computer games and video surveillance, have become more and more popular with the rapid development of computer technology. There are many approaches to create a virtual human body. Common ways include employing 3D body scanners or using 3D modeling techniques based on understandings of human anatomy. Detailed 3D human body model can be acquired easily using such methods. However, 3D body scanning is expensive and clumsy, and more seriously, the person to be modeled might not be available for scanning. Meanwhile, human models based on human anatomy generally fall short in representing personalized individuals.

SHORT COMMUNICATIONS proceedings
ISBN 80-86943-05-4
WSCG'2006, January 30-February 3, 2006
Plzen, Czech Republic.

Another approach recovers the human model using image(s). This approach can be divided into two groups: (1) reconstruction from static human figure and (2) reconstruction from human figure in motion.

Most researchers in the first group employ multi-view images of a human figure in a specific posture [Hil98a; Lee00a; Vil03a]. There are strict requirements on the images which must capture certain specific views of the immobile human figure. The human figure is required to appear in certain standard immobile postures. Body reconstruction from image sequences containing human motion is another group in image-based human model reconstruction. Some researchers require the human figure to perform specified motions [Che03a; Kak95a] or limited arbitrary motions [Coh02a; Dap00a; Sta03a] under multiple cameras. No matter whether the static human figure or human motion is used, methods using multiple cameras share the same drawback: the person has to pose for the cameras at a specific location, normally in a fully-equipped laboratory or studio. In contrary, monocular video is conveniently available in various formats (e.g. DVD, VHS video tape, mpeg). However, monocular video is the most difficult source for body reconstruction, especially when the camera is uncalibrated which is mostly the case. Most reconstruction work based on monocular images ([Bar03a; Rem03a; Tay00a]) use the so-called "scaled-orthographic" camera model, which amounts to a parallel projection, with a scale parameter added to mimic the effect that the image of the object shrinks with the distance. Obviously the scaled-orthographic camera model is very different from the real camera used to capture the video. The only exception is [Pen05a] who used a perspective camera model to handle the monocular images. Among these research effort, ([Bar03a], [Pen05a]) tried to reconstruct the human skeleton from uncalibrated monocular images, while others ([Rem03a; Tay00a]) attempted to reconstruct the human motions with a pre-defined human skeleton model.

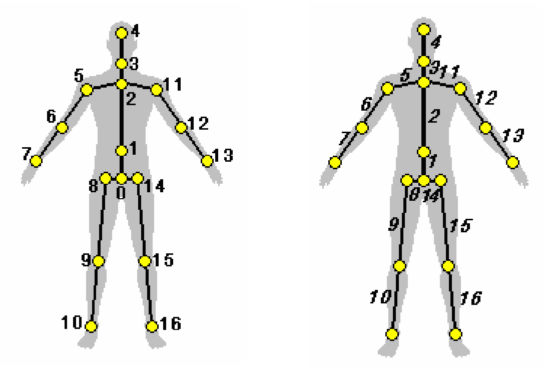
Human body reconstruction from monocular image sequence is a very attractive idea due to the fact that monocular video is so conveniently available. For the appearance of the whole human body, the skeleton proportion is a dominant feature. Recovering the proper skeleton model is therefore the most important component for the reconstruction of human body. Due to the vast possibilities of human postures and camera settings, skeleton proportions usually cannot be simply calculated from the projected segments. To estimate such proportion, [Bar03a] requires the user to manually identify the body parts almost parallel to the image plane, or the skeleton segments having similar orientation with respect to the camera. From these manual identifications and under the scaled orthographic model, their algorithms are able to choose a suitable stick model from a group of pre-defined models. It is very troublesome to identify the skeleton segments with certain properties. When the input source contains large number of frames, such identification becomes impossible. Further more, analysis without using real perspective camera model results in serious estimation error especially when the perspective effect is significant in the images. [Pen05a] proposed a method to extract the human skeleton model from a perfect synthesized data sequence with a perspective camera model. They assume the input data are perfect without any noise and the camera is fixed with no motions.

It is clearly desirable to develop algorithms to estimate the skeleton proportions from only the landmark joints using a real perspective camera model which is more flexible. In this paper, we propose a novel method to extract the human skeleton proportion from uncalibrated monocular data. The human motion is unrestricted and a perspective camera model is used. Superior to [Pen05a], the proposed method allows the source data captured by the camera performing some simple motions, e.g., rotating, zooming.

The rest of this paper is organized as follows: Section 2 provides the overview of the system; Sections 3 and 4 discuss the major parts of the proposed system; Section 5 demonstrates the results with discussion; and Section 6 concludes the paper.

2. SYSTEM OVERVIEW

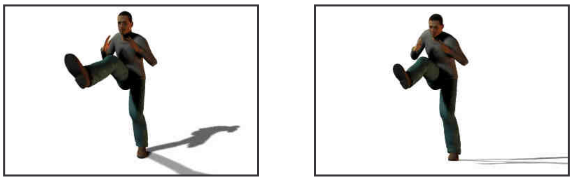
Human skeleton system can be treated as skeleton segments connected at joints as shown in Figure 1. It can be represented hierarchically as a skeleton tree if one joint is taken as the root. To minimize the number of tree levels, the joint J1 is considered as the root joint. It is clearly impossible to derive the absolute lengths of any body segments from monocular images without the knowledge on camera setting and the distance of the human to camera. Hence the skeleton proportion of every skeleton segment is used to represent the skeleton model which is defined as the ratio of the individual length of every skeleton segment to the sum of the lengths of all segments.



(a) Joint IDs (b) Segment IDs

Figure 1. Skeleton joint IDs and segment IDs

Due to the considerable diversity of camera types and media types, the images recording the real world may be different in image size and aspect ratio. The conventional film is 35mm film with the size of 36mm×24mm. Camera parameters based on 35mm film, the basic film format, are commonly used as the standard. Therefore, without much generality loss, we scale any input image based on its original aspect ratio to fit and centering into a 35mm film for normalization.



(a) strong (b) little

Figure 2. Perspective effects

To reconstruct the human skeleton from uncalibrated monocular images, [Bar03a] uses a scaled-orthographic camera model. In that work, the user is required to identify the perspective effects in the image(s), and only those with very little perspective effect can be chosen as the input. Such limitation greatly reduces the range of usable input video. In contrary, the system proposed here accepts input video with any levels of perspective effect. The user only needs to give an initial focal length to the system by identifying the perspective effects in the first few images. A reference guideline is provided for strong to little perspective effects, indicating the focal length range from 20mm (wide angle) to 300mm (telephoto). Figure 2(a) gives an example image displaying the strong perspective effect while Figure 2(b) shows an example with little perspective effect. If the user is unable to determine the perspective effects in the first few images, a default initial focal length is given as 50mm, which has the perspective effects closest to those seen by normal human eyes [Ele05a].

Many video available are recorded by a mounted camera on a stationary tripod. The mounted camera may be static or performing certain motions such as zooming, panning etc. during the video capturing. The proposed system estimates the camera motions from the source input data and “undo” these motions to produce the modified 2D data. The modified 2D data can then be considered as always taken by a static camera.

A human skeleton modeling algorithm is then used to estimate the proportions of the human skeleton from the 2D monocular data taken by a static camera.

3. CAMERA MOTION AND STANCE FOOT

Before applying the skeleton modeling algorithm for 2D monocular data taken by a static camera, a system is proposed to estimate motions such as zooming, panning etc, of a mounted camera. Such information is then used to “undo” the camera motions to produce the modified 2D monocular data as if they are taken by a static camera.

To estimate the camera motions from one frame to another, at least 3 pairs of corresponding points are required [Tan95a]. Each pair of corresponding points must be projected from the same stationary point. In fact, about 10 such pairs are needed for satisfactory estimation results [Tan95a]. Therefore, if the image background is featureless and no assistant stationary object is available, it is almost impossible to estimate the camera motions between two frames, since it is hard to find even one pair of stationary points.

In this project, the only input information is the projections of the 17 skeleton joints. As the human is dynamic, none of these skeleton joints can be considered stationary at any time. Without a pair of corresponding points between two frames, it is impossible to calculate the camera motion. Fortunately in most human motion, there is always a foot to support the weight of the body at any moment. This foot is called the stance foot and usually remains at the same position for a short period of time. We make use of this small piece of information to estimate the camera motions.

3.1 2D Data Rectification

Human motions and camera motions usually exhibit certain coherence in time. Thus, the projection trajectory of each human skeleton joint within a few neighboring frames is usually smooth. Any inconsistencies are deemed to be brought in by the sampling error in data extraction. Therefore, the 2D data are first rectified by a smoothing filter.

3.2 Sequence Segmentation

When recording video with a mounted camera, the camera can either be static (motionless) or perform the following motions: (1) zooming: the camera's focal length changes; (2) panning: the camera rotates about the Y-axis; (3) tilting: the camera rotates about X-axis; (4) swinging: the camera rotates about the Z-axis. (See Figure 3)

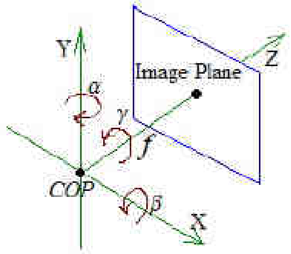


Figure 3. Camera motions

In practice, a mounted camera often remains static, occasionally performing one or both of the following motions: (1) directional rotating: a combination of panning and tilting in one direction, usually used when tracking a moving person; (2) zooming: usually used when keeping the full body of a person in the film range, if the person moves closer or further to the camera. This project assumes the camera performs only one of these motions at a time, and the camera would be motionless for at least 1 second in between any change of its motion. It is also assumed that the duration of each camera motion is over 0.5 second. On the other hand, each stance foot can be reasonably assumed to be static for more than 1 second. Under these assumptions, it is possible to find the frames when there is no camera motion, since the projection of the stance foot would be fixed during those periods.

For the input data with frame rate *p* fps, the projection of a stance foot would remain fixed for at least (*p/2*) frames under the assumptions that the camera remains still for at least 1 second before any motion and the stance foot remains still for at least 1 second. The extreme case is that, within 1.5 seconds, the camera moves during the first 0.5 second and remains still for the next 1 second, while the foot is only still for the first 1 second. In this case, the projection of the foot is static for the period from 0.5 to 1 second, i.e. *p/2* to *p* frames. Therefore, the accumulated value of the movement of each foot in every neighboring *p/2* frames can be calculated. If the accumulated movement of any one foot in these frames is below a given threshold, these frames are considered to be taken by a static camera.

In this way, the input 2D data is segmented into different sequence segments. These segments can be classified into two categories: data from static camera, and data from camera in motion.

3.3 Camera Motion Estimation

For video segments that come from dynamic camera, camera motion needs to be estimated in order to produce the modified 2D data. If the stance foot is known in each frame within the data segment, the camera motion estimation is pretty easy. However, most of the time it is not known which foot is the stance foot in these frames. An iteration method is hence proposed to determine the stance foot in each frame.

It is assumed that the stance foot usually remains motionless for at least 1 second. For the input with frame rate *p* fps, the stance foot would not lift off within *p* frames after it first touches the ground. Therefore, for the video segment of *n* frames, there are minimum *0* time and maximum (*n/p+1*) times of stance foot swapping.

Every possible case of foot swapping is tested to calculate the motion smoothness under the two possible camera motions mentioned in the previous subsection. The “motion smoothness” is calculated based on the standard deviations in the motion parameters: For “directional rotating”, the motion parameters are the panning and tilting angles between two neighboring frames; for “zooming”, it is the scaling difference in focal lengths between two neighboring frames. The smoother a camera motion is, the more likely that motion is correct.

After testing all the cases, two possible cases can be discovered: one with the largest motion smoothness for “directional rotating”, another with the largest motion smoothness for “zooming”. The case with the large (above the threshold) motion smoothness is chosen as the solution. If both motion-smoothness are large (above the threshold), both camera motions are possible. In this case, the best way is to “undo” the camera motions in both cases. Significant enlargement/shrinking effects will appear in the new data if the camera motion is “undone” mistakenly and the right camera motion can be chosen correspondingly.

3.4 Modified 2D Data Production

After the camera motions are estimated for each segment from the dynamic camera, it is possible to “undo” the camera motions for the whole input 2D data segment by segment. The resulted modified data can be treated as taken by a static camera. This static camera also has the initial focal length chosen by the user.

4. SKELETON MODELING

The skeleton modeling process is based on the 2D data from static camera and the initial focal length selected by the user. It consists of two major steps: (1) virtual scale parameter estimation; (2) length of skeleton segment calculation.

4.1 Virtual Scale Parameter

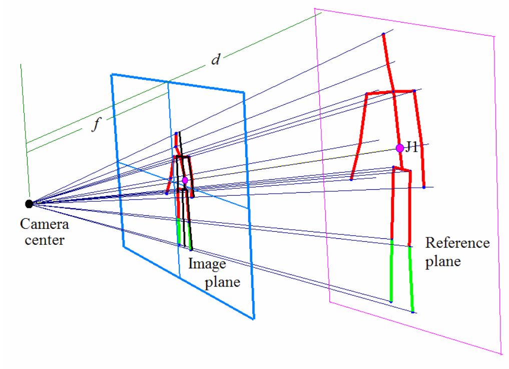


Figure 4. Scaled human figure

Any image can be considered as the projection of a virtual human figure whose root joint lies on the image plane. As shown in Figure 4, if the plane passing through the root joint J1 of the actual human figure has the distance *d* to the camera center, the scaling parameter of the human on the image can be calculated as $s=f/d$, where *f* is the camera’s focal length. However, the scaling parameter in each frame is impossible to establish since the actual distance *d* is unknown. Assume that such distance for every frame in the sequence are d_1, d_2, … , d_n respectively. Assuming that the first frame has the unit scaling parameter ($d_1'=f$), we just need to find the relative distance to derive d_1', d_2', …, d_n'. If the distance relationships are correctly found, the distances d' calculated would have the same trend as the actual d: $d_1'/d_1= d_2'/d_2=\ldots= d_n'/d_n$. The virtual scale parameter is taken as $s'=f/d'$, with the first frame having $s_1'=1$.

To find the distance relationship for the joint J1, the standard perspective projection theory can be utilized. It is based on the fact that the projected length of any line segment parallel to the image plane has a linear relationship with its distance from the image plane - the further the segment is from the image plane, the shorter its projected length is. Obviously such knowledge cannot be directly used for joint J1, since it does not necessarily belong to a line segment parallel to the image plane.

For this purpose, the projected length of the calf connected to the stance foot is studied. It is difficult to gauge the posture of the actual calf only from its projection. However, the frame containing the longest projection of that calf can be treated as taken at the moment when the calf is parallel to the image plane. Searching through the entire sequence, the projected lengths of the possible parallel calf from different stance foot can be obtained from the input.

Such information can then be used to derive the distance relationship of the joint J1. During the swapping of the stance foot, the joint J1 is usually located on the same plane that passes through the two feet and is perpendicular to the ground plane. If a reference plane passing through J1 is parallel to the image plane, any point on the intersection line between this plane and the ground can be utilized for calculating the virtual distance of J1. One straightforward way to approximate such a point is to find the intersection between the vertical line passing through J1 and the line segment from one stance foot to the other. With this point, the virtual distance in this frame can be calculated. However, not all frames show the moment of stance foot swapping. The virtual distance in other frames are then interpolated linearly. The virtual scale parameter can then be derived from these virtual distances.

4.2 Virtual Length Estimation

Figure 5 shows the projection of a branch of the connected skeleton segments from the root joint. The virtual scale parameter has been derived for this image, and the image is captured by a camera with a known focal length *f*. P_1' is the projection of the root joint; line segment P_i'P_{i+1}' represents the projection of the skeleton in the i^{th} level of the skeleton tree; and the projection center is denoted as O_C. It is assumed that this frame is projected from a virtual human figure whose root joint P_1 locates on the image plane, as discussed in "Virtual Scale Parameter" part.

The estimation of the length of the virtual skeleton is preceded hierarchically. The estimation in the first level of the tree includes two steps: (1) lower boundary estimation for all possible virtual skeleton length in each frame and (2) virtual skeleton length selection.

As shown in Figure 6(a), P_1'P_2' is the projection of a skeleton segment in the first level of the tree while the root joint P_1 of the virtual human figure is located on the image plane. P_1'P_2' can be projected from an infinite number of possible skeleton segments with different lengths, due to the uncertainty of the actual joint that projects onto P_2'.

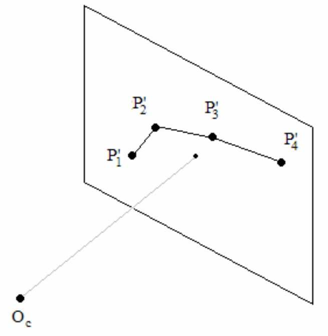
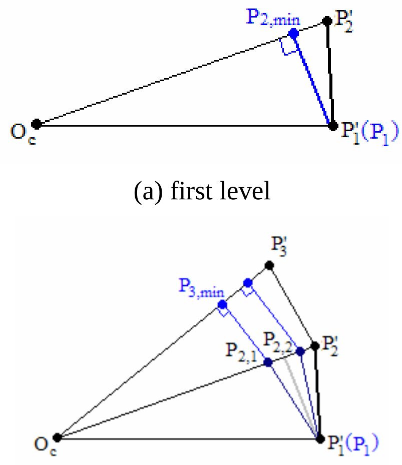


Figure 5. Projection of connected segments



(a) first level

(b) second level

Figure 6. Estimation of lower boundary

Fortunately, the lower boundary of these possible skeleton lengths can be calculated, e.g., $P_1P_{2,min}\perp O_CP_2$', where $|P_1P_{2,min}|$ is the shortest possible virtual skeleton length with end point P_1 that has the projection P_1'P_2'. This shortest possible virtual length will always be shorter than the actual virtual skeleton length in this frame, and is called the lower boundary for this frame.

With the virtual scale parameter in every frame established, the lower boundaries from all frames are comparable. Denote the virtual scale parameter in all frames as: s_1', s_2', …, s_n' , the lower boundary of the virtual skeleton length in each frame as: l_1, l_2, … l_n respectively, and the skeleton length of the scaled human figure at the normalized size as L. According to the calculation of lower boundary in each frame, it is known that $l_i/s_i' \leq L$ ($1 \leq i \leq n$). Therefore, the largest value of these normalized lower boundaries in all frames $Max\{l_i/s_i', 1 \leq i \leq n\}$ is the best approximation of the virtual length L of the actual human at normalized size.

Unlike the segments in the first level where one end joint is assumed to be on the image plane, both end joints of the segments in higher levels are undetermined. Hence, the position of one end joint must be estimated first in order to calculate the shortest possible virtual length. If the virtual lengths of the segments in the lower levels of the skeleton tree are known, there are at most 2^i possible skeleton poses in the i^{th} level. For example in Figure 6(b), $P_{2,1}$ and $P_{2,2}$ are two possible positions that give the projection at P_2' in 2nd level. As a result, 2^i shortest possible relative lengths in level i^{th} can be calculated. The minimum value among them is selected as the shortest possible virtual length of this skeleton in this frame. Similar to the case in the first level, the largest value of the normalized shortest possible virtual lengths among all frames has the minimum difference from the actual length of that segment at the normalized size.

In this way, the skeleton length for each segment can be estimated under the given focal length.

5. RESULTS

One of the most challenging tasks in any 3D reconstruction attempts from images is the 2D feature extraction from images. Up to date, the image processing techniques are still unable to provide sufficiently accurate 2D feature information from any images. In fact the extraction of 2D feature information and the 3D model reconstruction are two independent modules. They should be addressed separately and in parallel for the ultimate development of the complete system. In this project, the image processing step is not an emphasis. 2D extracted projections of human skeleton joints are assumed available as the input of our system.

Computer synthesized input video is currently used to test our algorithm. The synthesized video is generated by computer with a 3D virtual skeleton model and a virtual camera. The 3D virtual skeleton model has 17 joints with 16 skeleton segments. It is driven by BVH motion files [Ani05a]. The virtual camera is located at a fixed 3D position when capturing any frame of 720*480 pixels, where the fixed 3D position of the camera is not restricted as long as the camera can capture the whole human motion. It can perform one of the following motions in between motionless periods: directional rotating (panning and/or tilting, but excluding swinging since it is hardly used in video production) and zooming. Each camera motion, including the motionless period, should be smooth and lasts for at least 0.5 second. Noises are added to the 2D data to make the application closer to actual practice.

The proposed system first estimates the camera motions in the source 2D data. The modified 2D data is then produced which can be considered as taken by a static camera. The human skeleton model in different frame has different projected size. A virtual scale parameter is needed to resize the human model to a standard size. The supporting foot determined in each frame can help to estimate the virtual scale parameter. After that, the length of the same skeleton segment in each frame is estimated. By using the virtual scale parameter, these lengths can be compared and the length for that skeleton segment can be determined for the human model at the standard size.

5.1 Camera Motion

To estimate the camera motion, the data segments from a dynamic camera are detected from the source data. Two possible camera motions can be estimated: “directional rotating” and “zooming”.

Figure 7 shows some sample frames comparing the source data and the modified data. The source data includes three camera motions: panning, tilting and zooming. With the initial focal length specified by the user, the proposed system successfully identified the data segments that taken by a dynamic camera. The camera motion for each of such segments is also automatically determined.

5.2 Virtual Scale Parameter

The proposed system utilized the estimations of stance foot and the recovered 2D data to calculate the relative distance for the position of each stance foot. With this relative distance, the relative distance of joint J1 in each frame is determined. The virtual scale parameter is estimated based on this relative distance.

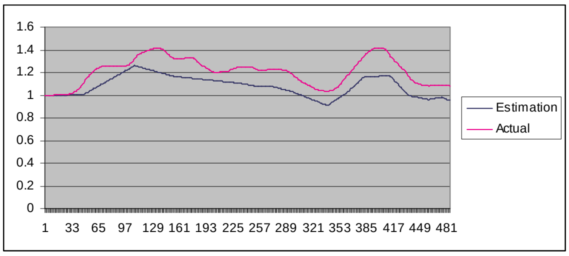


Figure 8. Virtual scale parameter estimation

Figure 8 shows the result comparing the estimated virtual scale parameter to the actual one. The vertical axis represents the virtual scale parameter while the horizontal axis represents the frame id. The red dots indicate the actual virtual scale parameter at each frame while the blue dots indicate the estimation. From the comparisons, it can be seen that the trend of scaling up and down can be successfully estimated.

5.3 Skeleton Model

The estimation of the skeleton proportion is shown in Table 1 which compares the actual and the estimated skeleton proportion in both numerical and graphical ways. The first column indicates the skeleton segment ID. The second column and third column show the skeleton proportion of that skeleton segment. The skeleton proportion is calculated by comparing its length to the sum of all skeleton lengths. The fourth column shows the difference between the estimated and the actual proportion. The last column compares the skeleton proportion in an intuitive way: the red skeleton with blue joints is the actual skeleton model while the green skeleton with black joints shows the estimation.

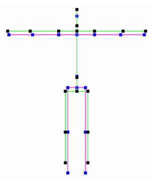
Skeleton	Actual	Estimation	Error	Comparison
1	4.90%	5.97%	1.07%	
2	17.75%	19.86%	2.11%	
3	1.55%	2.28%	0.73%	
4	6.44%	6.89%	0.44%	
5	7.54%	7.69%	0.15%	
6	11.19%	12.17%	0.98%	
7	10.08%	9.58%	0.50%	
8	3.80%	4.66%	0.86%	
9	19.05%	17.50%	1.55%	
10	17.69%	13.40%	4.29%	

Table 1. Skeleton proportion estimation

It can be easily seen from Table 1 that the estimated skeleton proportion resembles the actual skeleton proportion very closely. The errors between the estimation and the actual skeleton model are very small in lower hierarchical levels but increase in higher levels. In skeleton S10, the estimated error reaches its peak. The estimation accuracy of the skeleton in higher level decreases due to the algorithm's hierarchical nature.

6. CONCLUSION

This paper presents a simple but effective method to estimate the human skeleton proportion from uncalibrated monocular data. The proposed method can estimate the human skeleton proportion without using any pre-defined skeleton information. It is not required to manually identify skeletal segments with certain properties, and no posture estimations are necessary. The proposed method is tested on the computer synthesized data captured by a mounted camera with motions. The reconstructed result is satisfactory. Future work includes the improvement on the estimation of skeletons in higher hierarchical levels. More cues are to be investigated in order to minimize the error propagated to higher hierarchical levels. The algorithm will also be tested on real video to verify its applicability.

7. REFERENCES

[Ani05a] Animazoo.BVH Library. http://www.bvhfiles.com, accessed 1 Jul, 2005.

[Bar03a] Barron, C., and Kakadiaris, I.A. On the improvement of anthropometry and pose estimation from a single uncalibrated image. Mach. Vision Appl., 14(4):229-236. 2003.

[Che03a] Cheung, K. M., Baker, S., and Kanade, T. Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture. CVPR. 2003.

[Coh02a] Cohen, I., and Lee, M. W. 3D Body Reconstruction for Immersive Interaction. 2nd International Workshop on Articulated Motion and Deformable Objects. 2002.

[Dap00a] D'Apuzzo, N., Gruen, A., Plankers, R., and Fua, P. Least Squares Matching Tracking Algorithm for Human Body Modeling. XIX ISPRS Congress, Amsterdam, Netherlands. 2000.

[Ele05a] Elert, G. (editor). Focal Length of a Camera Lens, The Physics Factbook. http://hypertextbook.com, accessed 1 Jul, 2005.

[Hil98a] Hilton, A., and Gentils, T. Popup people: capturing human models to populate virtual worlds. SIGGRAPH. 1998.

[Kak95a] Kakadiaris, I. A.,and Metaxas, D. 3D human body model acquisition from multiple views. ICCV, pp. 618. 1995.

[Lee00a] Lee, W-S., Gu, J., and Magnenat-Thalmann, N. Generating Animatable 3D Virtual Humans from Photographs. Computer Graphics Forum (Eurographics 2000), 19(3). 2000.

[Pen05a] Peng, E., and Li, L. Estimation of Human Skeleton Proportion from 2D Uncalibrated Monocular Data. CASA. 2005.

[Rem03a] Remondino, F., and Roditakis, A. Human figure reconstruction and modeling from single image or monocular video sequence. 3DIM 2003, pp. 116-123. 2003.

[Sta03a] Starck, J., and Hilton, A. Model-Based Multiple View Reconstruction of People. ICCV, pp.915. 2003.

[Tan95a] Tan, Y-P., Kulkarni, S.R., and Ramadge, P.J. A New Method for Camera Motion Parameter Estimation, ICIP. 1995.

[Tay00a] Taylor, C. J. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. Comput. Vis. Image Underst., 80(3):349-363. 2000.

[Vil03a] Villa-Uriol, M., Sainz, M., Kuester, F., and Bagherzadeh, N. Automatic creation of three-dimensional avatars. Videometrics VII, Proceedings of the SPIE, 5013:14-25. 2003.

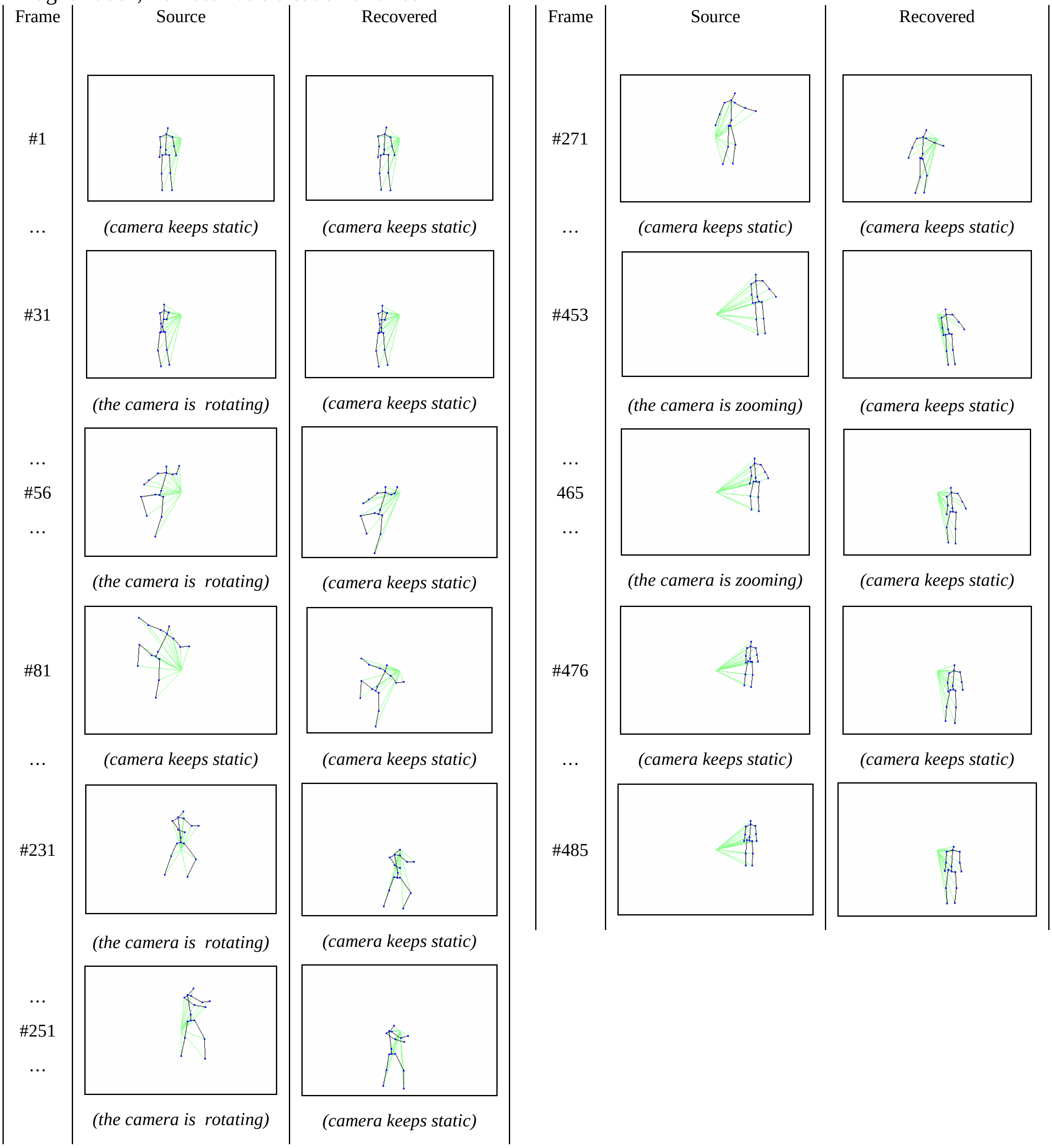


Figure 7. Camera motions and the modified 2D data