Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Robotics and Automation, Madrid, Spain, February 15-17, 2006 (pp381-386)

Human Skeleton Proportions Recovery from Monocular Data

EN PENG LING LI
Department of Computing
Curtin University of Technology
GPO Box U1987 Perth, WA 6845
AUSTRALIA

Abstract: - This paper introduces a novel method for recovering the skeleton proportions for a human figure from monocular data. A perspective camera model is defined semi-automatically based on the user's judgment. In the proposed method, key frames are first extracted from the source data automatically. A human skeleton model is then constructed to match all key frames under the established camera. An advantage of the proposed method is that no human posture validation is needed during the modeling process. The proposed method is tested to propose satisfactory results for some input data. The human model with recovered proportions can be used in further research involving body reconstruction or human motion reconstruction.

Key-Words: - Modeling, human figure, monocular

1 Introduction

The problem of creating a virtual human has received a lot of attention for the last few decades, due to the increasing popularity of applications involving human figure, such as movies and computer games. There are many approaches to create a virtual human body.

Traditional ways for creating a virtual human figure include employing the 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, it is expensive and clumsy to use the scanner, and more seriously, the person to be modeled might not be available for scanning. Meanwhile, the human models based on human anatomy generally fall short in representing personalized individuals. Another way recovers the human body using images, the popular media that can record the human figure.

The research using images can be divided into two groups: (1) using multi-view images; and (2) using single-view images. Multi-view images are adopted by many researchers: Some researchers [6] [8] [14] recover the human body from multi-view images recording a static human figure; some researchers reconstruct the virtual human from multi-view images recording a dynamic person performing specified motions [3] [4] [5] [7] [12]. 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, single-view images are conveniently available in various formats to general public. Hence, human body reconstruction from single-view images is a very attractive idea.

The approaches to recover the human figure from single-view images can be separated into two groups depending on the camera model adopted: (1) using affine camera model; and (2) using perspective camera model. Affine camera model is the approximation of the real camera model, which has three important instances: orthographic model, weak perspective (a.k.a. scaled-orthographic) model, and paraperspective model. Among these instances, weak perspective camera model is popularly used by the researchers [1] [11] due to its simplicity. However, the weak perspective camera model can only handle the images with very little perspective effects, since the affine camera model does not represent a real camera. To handle the images with any perspective effects, perspective camera model is required since it represents a real camera.

Due to the complexity of using perspective camera model, there are limited research efforts [9] [16]. Some researchers [16] restrict all body segments of the human figure as almost parallel to the image plane in order to acquire accurate human skeleton proportions. Some researchers [9] require estimating the virtual scale parameters for each

frame. Such estimation may lead to large inaccuracy.

It is clearly desirable to develop new algorithms for recovering the virtual human figure from single-view images using a perspective camera model. A human skeleton proportions estimation system with such new algorithms is proposed in this research.

The limitations for the proposed system include: (1) the camera is fixed and parallel to the ground during capturing; and (2) at any moment during capturing, there is at least one foot touches the ground.

Despite these limitations, there are many merits in the proposed system. The major advantages include: (1) the perspective camera model is used - which enables the ability of handling images with any perspective effects; (2) the input source is not so restricted - the human figure in the input source does not require to be almost parallel to the image plane; (3) only the visible feature points are utilized by the proposed system - unlike [9] [16] which assume all feature points are available even if the feature points are occluded.

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

2 Overview

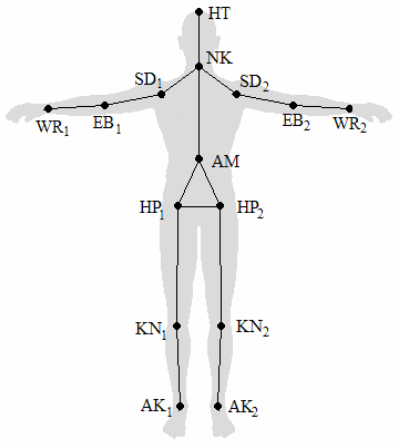


Figure 1. Feature points of human figure

Human figure is formed by skin, muscle and skeleton. The skeleton dominates the appearance of a human figure. The human skeleton is a number of skeleton segments connected at skeletal joints. The most important feature for the skeleton is the proportion, which is regardless with the height of the human figure. The skeleton proportions for different people may vary. 15 feature points are made of which are assumed available from any input source as long as they are visible, as shown in Figure 1.

The focal length of the perspective camera model can be semi-automatically determined by the user [9]: the user judges the level of perspective effects from the input source, and then the algorithm automatically maps the perspective level to corresponding focal length. A virtual scene can be constructed for each image by assigning the virtual film and a virtual ground. The virtual film is in 35mm film format. Any image can fit into this film by scaling without changing the original aspect ratio. The virtual ground is assumed parallel to the camera's orientation and passing through the bottom of film. Based on the basic knowledge of perspective projection, if there is a virtual human figure with the same proportions as the original one in the source image, it is possible to reconstruct the posture on the virtual human which is identical to the real scene during capturing the image. The proposed research will recover the skeleton proportions for the virtual human figure. The size of the virtual human figure in each virtual scene should be identical. The algorithm for recovering the human skeleton proportions will be discussed in Section 4. The virtual human figure with recovered human skeleton proportions is actually a scaled version of the real human figure.

The main human proportions recovery algorithm works mainly on the frames where both feet are on the floor. Such frames are defined as key frames in this research. The recovery process consists of two components: (1) extraction of key frames; (2) human model acquisition from key frames.

3 Key frames extraction

The input 2D data contains a dynamic human figure with at least one foot placed on the floor at any moment. Therefore, each input frame displays either both feet touch the floor or only one foot touches the floor.

Since the camera is mounted at a fixed position, if the projection of a foot remains at the same 2D position within a few neighboring frames, there are two possibilities: this foot remains at the same 3D position during this time, or this foot moves along the straight line passing through the camera center

and the foot itself. Such foot can be considered as a candidate for motionless foot. Hence, the candidates for motionless foot can be extracted by observing their projections within a few neighboring frames.

However, due to errors in feature extraction, such as noises, the extracted feature point of a static foot may not accurately represent its actual projection. It is possible that both feet "appear" to be dynamic, if only judging from their projections. Therefore, to test whether a foot is static in a frame, its deviation in a few neighboring frames are added up. If the accumulated deviation is over a threshold value, that foot is considered dynamic, otherwise it is considered static. This way, it is possible to determine whether a feature point in one frame is the projection of a static foot.

After the static foot is found, it is necessary to determine whether it is on the ground. The algorithm will automatically find the frames indicating foot swapping, an important interaction between the feet and the ground. From such information, it is not difficult to eliminate those frames that may include the static foot in the air. If any frame can be represented by the foot status: "L"-left foot is static, "R"-right foot is static, "B"-both feet are static, "N"- none of the feet is static. For example, there are ten neighboring frames represented as "LLLBBBLLLL". In the first 3 frames and last 4 frames, left foot must be a stance foot because only it is static; but for the 4^{th}, 5^{th} and 6^{th} frame, both feet is static, any foot can be the stance foot. However, only the left foot can be guarantee as the stance foot in these three frames, because left foot remains static for these frames. Instead, the right foot may be stance foot or a static foot in the air. Then, the status for these ten frames can be updated as "LLLLLLLLLL". There are many circumstances that require further determination. But the focus of this paper is the skeleton proportion estimation algorithm. Hence, the discussions on other circumstances are skipped.

From the analysis of foot positioning, key frames can be automatically selected by choosing the frame with both feet on the ground.

4 Basic estimation system

The basic idea for this component is to find a possible human model which can reproduce the projection data under the virtual camera for all key frames.

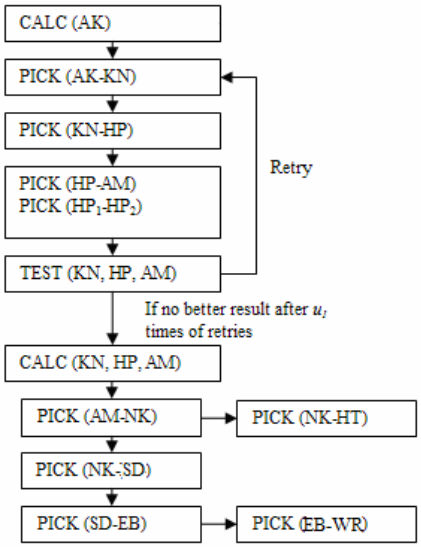


Figure 2. Basic estimation system

It is well-known that any projection point on the viewing plane can be back-projected to infinite number of possible 3D positions. Fortunately, key frames containing the human with both feet on the ground have been selected. Therefore, the projections of both feet in such frames can be known as projected from a point on the virtual ground. With the perspective camera model, the 3D positions of both feet in each virtual scene are unique and can be easily calculated.

Starting from the 3D position of both feet in the virtual environment, the lengths of each skeleton segments of the human body are recovered according to the following scheme, separated as lower body part and upper body part respectively.

The estimation process is shown in Figure 2. During the estimation process, there are three basic actions "PICK", "TEST" and "CALC". They are implemented as below:

4.1 Action "PICK"

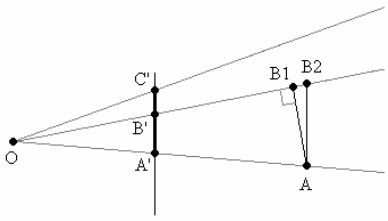
The purpose of this action is to assign a length for a specified skeleton segment. This length should be within the range of all possible skeleton lengths. Hence, the lower boundary and the upper boundary of the range should be firstly determined.

The lower boundary of the range is defined as follows: If the skeleton segment has the length below this lower boundary, it is not possible to have its projection matching all key frames under the estimated camera model from the given 3D position. Instead, if a skeleton is longer than the lower boundary of the range, it will forever be able to produce the given projection.

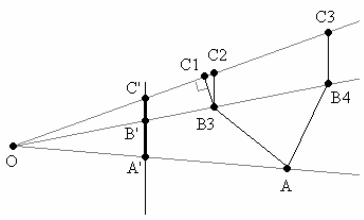
The upper boundary of the range is introduced for the purpose of minimizing the searching range. It is defined to allow a skeleton segment to give the projection from any possible 3D position.

To acquire the lower boundary and upper boundary, the shortest possible length and longest possible length for each key frame must be calculated.

An example from one frame is shown in Figure 3 where point O represents the camera center, points A', B', and C' are the projections on i^{th} frame. In Figure 3(a), the 3D position of point A for this scene is known. In order to have the projection A'B', skeleton segment AB will never be shorter than $|AB_1|$. Thus $|AB_1|$ is the shortest possible length for skeleton segment AB in this frame. $|AB_2|$ is the longest possible length defined for skeleton segment AB in this frame, since there is only one possible 3D position for point A. Similarly, in Figure 3(b), the 3D position of point A and the skeleton length of AB are known. In order to have the projection B'C', the skeleton segment BC should not be shorter than $|B_3C_1|$ under every possible 3D position of point B. So, $|B_3C_1|$ is the shortest possible length for skeleton segment BC in this frame. $|B_4C_3|$is the longest possible length defined for skeleton segment BC in this frame.



(a) One level



(b) Two levels

Figure 3. Shortest possible and longest possible length of a skeleton segment

Denoting the shortest possible length of a skeleton segment's length in i^{th} key frame (total n key frames) as $L_{min}(i)$ and the longest possible length as $L_{max}(i)$, the lower boundary and upper boundary of the range $[L_{min}, L_{max}]$ for the skeleton's segment for all frames can be calculated as:

$$L_{min} = \text{Max}\{ L_{min}(i) \mid i=1,2,\ldots,n \}$$

$$L_{max} = \text{Max}\{ L_{max}(i) \mid i=1,2,\ldots,n \}$$

Any length within this range is possible for the skeleton segment. To make the algorithm more efficient, the length is randomly picked from this range and assigned to the skeleton segment.

4.2 Action "TEST"

The purpose of this action is to test whether the feature points can be put to the reasonable positions. There is one such action in the proposed algorithm: the test is performed on feature points KN, HP and AM.

When testing KN, HP and AM, the skeleton length for segment AK-KN, KN-HP and HP-AM are already estimated. For the projection of each leg, since the 3D position of AK is known, there will be maximum 2 possible positions for KN and maximum 4 possible positions for HP under the perspective camera model. Then there will be at most 16 possible distances between joint HP in both legs. Each possible case will form a triangle HP1-AM-HP2. Among all frames, it is possible to establish a common triangle that is similar to the triangle of one of all possible cases in each frame. The difference between the common triangle and its most similar triangle in each frame is recorded to represent the estimation accuracy. Less difference represents higher accuracy of the estimation.

4.3 Action "CALC"

The purpose of this action is to calculate a unique 3D position of the feature points in the virtual scene for each key frame. For example, CALC (AK) will calculate the positions of the joint AK with the help of virtual camera and virtual ground. CALC (KN, HP, AM) will calculate the positions of joint KN, HP, and AM that can provide the best result in the action TEST (KN, HP, AM).

5 Results

The proposed system is tested on the data extracted from both synthesized video and real video. 2D feature extraction process is done in different way for synthesized video and real video: (1) for synthesized video, the projections of each joint can be calculated directly from the segmented silhouette in each image; (2) for the real video, the feature points are extracted manually at the moment.

5.1 Synthesized video

For the synthesized video, the segmented silhouette is extracted first for the automatic calculation of the feature points as shown in Figure 4.The feature points may not represent the actual projections of the corresponding joint due to noises or partly-occlusions.

(a) Original video (b) Segmented silhouette (c) Extracted feature points

Figure 4. Input data

A source video has been tested with our algorithm. This source video contains 485 frames of the human figure performing Kung Fu motion with the frame rate of 30fps, generated using POSER [10]. It is captured by a virtual perspective camera model with focal length of 20mm. The orientations of the camera is parallel to the ground during capturing the frames.

The focal length of the camera model is firstly estimated by the user's judgment [9]. Key frames are then automatically extracted based on the calculated 2D feature points as discussed in Section 3. The extraction results are illustrated in Figure 5. The blue columns indicate the actual key frames while the red columns indicate the estimated key frames. It can be seen that most extracted frames are key frames. Numerically, 96.5% of extracted frames in synthesized video are the actual key frames.

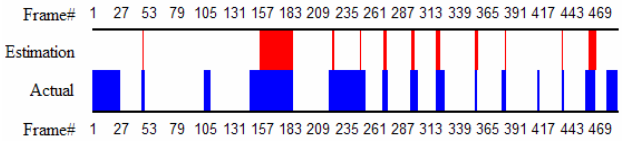
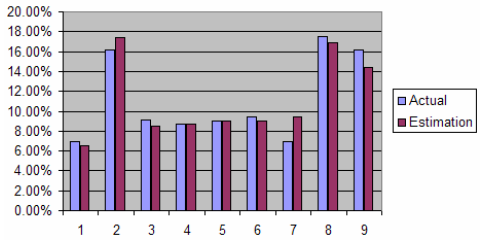
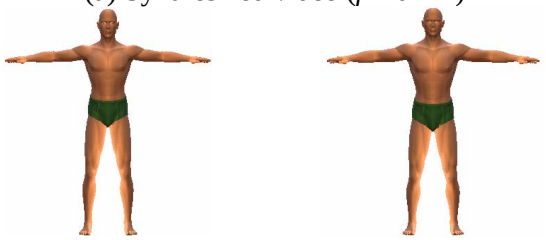


Figure 5. Synthesized Video (*f*=20mm)

The human proportion estimation system discussed in Section 4 is then applied on the extracted key frames. Figure 6 shows the estimation results. Figure 6(a) compares the proportion of each skeleton segment from the estimation and the actual ones by column chart while (b)~(c) compare them in intuitive form – the appearance of the human figure. In the column chart, *x* axis indicate the skeleton segment ID: 1(HP-AM), 2(AM-NK), 3(NK-HT), 4(NK-SD), 5(SD-EB), 6(EB-WR), 7(HP_1-HP_2), 8(HP-KN), and 9(KN-AK).



(a) Synthesized video (*f*=20mm)



(b) Actual (c) Estimation

Figure 6. Estimation results for synthesized video

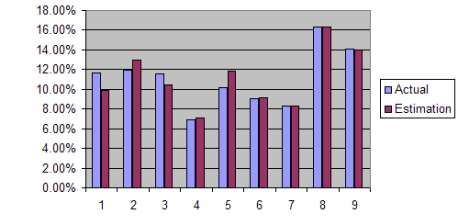
It can be seen from Figure 6(a) that the maximum error is about 2.5%. And the visual appearance of the estimated human figure is very close to the actual one.

5.2 Real video

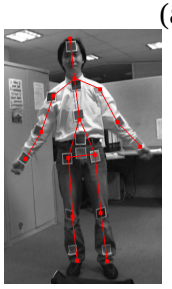
Feature point extraction from real video is a major problem in the research involving human figure. It is used or investigated by many researchers. They uses manual [1] [13], semi-automatic [5] or automatic [2] [15] algorithms. Manual labeling costs a lot of time. Semi-automatic and automatic algorithm cannot provide stable results. For this research, feature point extraction is not an emphasis, all feature points will be manual labeled.

Even if very short input real video contains a lot of frames, considering the normal frame rate 25fps. To avoid the heavy task for manual labelling, the key frames are manually selected. Although key frame extraction algorithm is not tested in real video at the moment, it is believed that key frame extraction algorithm will work well on real video if the feature points are available in all frames, provided the error is not very significant.

The human proportion estimation system is tested on the selected key frames from a real video. The camera is uncalibrated and is taken with a wide-angle lens. Figure 7 shows the results of estimation.



(a) Real Video



(b) Actual

(c) Estimation

Figure 7. Estimation results for real video

It can be seen from the chart in Figure 7(a) that the maximum error is 1.79%.

Based on the experiment results on synthesized video and real video, it can be concluded the proposed key frames extraction algorithm and the human proportion estimation system can produce satisfactory results.

6 Conclusion

This paper proposed a novel method to recover the human skeleton proportions from 2D uncalibrated data. The proposed method extracts the key frames, and estimate skeleton proportions of human figure. The proposed method is tested on both synthesized image data and real data. Both experiments achieve satisfactory results. The recovered human skeleton model is very close to the original and can be used in further research for full body reconstruction or motion reconstruction from monocular data.

References:

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

[2] J. Ben-Arie, D. Sivalingam and S. Rajaram, Probabilistic Labeling of Human Body Parts, *IASTED International Conference on Circuits, Signals and Systems,* Cancun, Mexico,pp.275-280 , May 2003.

[3] K. M. Cheung, S. Baker and T. Kanade, Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun, 2003.

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

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

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

[7] I. A. Kakadiaris and D. Metaxas, 3D human body model acquisition from multiple views, *ICCV '95: Proceedings of the Fifth International Conference on Computer Vision*, pp. 618, 1995.

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

[9] En Peng and Ling Li, Estimation of Human Skeleton Proportion from 2D Uncalibrated Monocular Data, *CASA*, 2005

[10] POSER, http://www.e-frontier.com/.

[11] F. Remondino and A. Roditakis, Human figure reconstruction and modeling from single image or monocular video sequence, *3-D Digital Imaging and Modeling 2003 (3DIM 2003), Proceedings, Fourth International Conference on*, pp. 116-123, 2003

[12] J.Starck and A.Hilton, Model-Based Multiple View Reconstruction of People, *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp.915, 2003

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

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

[15] Xiaofeng Ren, Alexander C. Berg, Jitendra Malik, Recovering Human Body Configurations Using Pairwise Constraints between Parts, *ICCV 2005*, 824-831, 2005.

[16] Jianhui Zhao, Human Animation from Motion Recognition, Analysis and Optimization, *Ph.D. Thesis, Nanyang Technological University*, Singapore, 2003.