A Model-based Approach for Human Motion Reconstruction from Monocular Images

Jianhui Zhao, Ling Li, and Kwoh Chee Keong

Abstract—In this paper, a model-based method is presented for human posture reconstruction and animation from un-calibrated monocular images that contain unrestricted human motions. A 3D skeletal human model is employed with encoded biomechanical constraints, and the model can be adjusted according to the 2D information of the extracted feature points. Energy Function is defined to represent the deviations between projection features and image features. Reconstruction and animation procedures are developed to translate or rotate joints and segments of the human model into their proper 3D positions with the help of Local Adjustment and Global Adjustment. This approach is a try to bridge the gap between computer vision and computer animation in human motion study.

Index Terms—Energy Function, Human Model, Monocular Images, Motion Recovery

I. INTRODUCTION

HUMAN posture recovery and animation from un-calibrated monocular images can be used in many fields [1]-[3], such as surveillance systems, rehabilitation, choreography, movie production, virtual reality, gesture recognition, game industry, etc. But it remains a challenging problem [4]-[5] due to the lack of information in the third dimension and the fact that the human body is an extremely complex object, being highly articulated and capable of a bewildering variety of motions. Projection from 3D to 2D is mathematically straightforward, but the inverse process is typically an ill-posed problem. Until now, there are only a few examples of 3D pose recovery from monocular images and most of them introduce simplifications such as movement restriction, application limitation, human interference, and non-robustness.

Chen and Lee [6] presented a method to determine the 3D postures of a human body from a walking film. Geometric projection theory is first applied to obtain a set of feasible postures, then dimensions of the human stick figure, physiological and motion-specific knowledge are used to constrain the postures in both single-frame and multi-frame analysis. But their work can only deal with human walking. Divide-and-conquer technique is another approach adopted by Holt et al. [7] for human gait and Ying Wu et al. [8] for human hand motion. Although it is attractive in terms of simplicity, this approach does not exploit the fact that all different components actually belong to the same object, and evaluating pose of each part separately may result in unfavorable combinations. M. Yamamoto and K. Yagishita proposed a method for tracking of a human body in 3D movement by using constraints from the scene [9]. Their approach utilizes constraints imposed on position, velocity and acceleration of every part of the body. These constraints can reduce DOFs of the model. This reduction guarantees the tracking problem to be a well-posed problem, and prevents tracking errors by noise. Thus the accuracy of tracking results will be improved if the subject moves under constraints in the scene. James M. Rehg et al. [10] developed a framework for figure tracking using a single video source. Their method of 3D reconstruction is then based on joint limits and key frame configurations. Joint limits reflect the limited range of moving skeletal joints, while manually specified 3D configuration of the model in selected key frames is used to improve the accuracy and resolve ambiguous solutions. C. Bregler and J. Malik [11]-[12] demonstrated a vision based motion capture technique that is able to recover articulated human body in video sequences. The visual tracking is based on an initialized first frame with known pose and angular configuration. 3D pose is found in minimizing the sum of squared differences between the projected model joint locations and the user supplied locations. The approach was tested by a person walking sequence in a front parallel plane. The human model is defined with a kinematics structure with one DOF in each of the joint. All joints can only move parallel to the Z-axis in the camera frame.

The novelty of our approach is that it is able to deal with any possible human motions from monocular images based on extracted feature points without user interaction or camera calibration in the reconstruction procedure. This paper is organized as follows: a 3D skeletal human model and its adjustment are introduced in section 2; Energy Function is defined in section 3; posture reconstruction and human animation procedures are described in section 4; experimental results are presented in section 5; finally, a brief conclusion is given in section 6.

II. HUMAN MODEL

In order to avoid constructing an over complicated 3D human model, we focus only on the major parts that participate in human motions. The basic idea is to regard the human body as an articulated object, which consists of 16 rigid parts
connected at 17 specific joints. As shown in Fig. 1, the joints are Hip, Abdomen, Chest, Neck, Head, Luparm, Ruparm, Llowarm, Rlowarm, Lhand, Rhand, Lthigh, Rthigh, Lshin, Rshin, Lfoot, Rfoot. The human body can be regarded as a tree structure at an abstract level. Each rigid part is allowed to move in its own domain with respect to its immediate ancestor.

![Fig. 1. Tree structure of the human model](image)

Kinematics analysis resolves any motion into one or more of six possible components: translation along and rotation about the three mutually perpendicular axes. By convention, these axes can result in adduction-abduction, flexion-extension and revolution motions (3 DOF). In our project, the skeletal human model is placed into an initial posture in the 3D space. Rotational ranges of every joint about X, Y and Z-axes of its local coordinate system based on biomechanical knowledge are used as geometrical constraints of human motion.

Geometric information of the adopted human model is not always consistent with that of the human subject in monocular images. Thus relative lengths of the body parts that are linked with \( p_{i,m} \) can be computed as:

\[
X_{image} = (p_{0,m}(x_0, y_0, z_0, k_0), \\
p_{1,m}(x_1, y_1, z_1, k_1), \ldots, \\
p_{n-1,m}(x_{n-1}, y_{n-1}, z_{n-1}, k_{n-1}); \\
x_i \in \alpha_i, y_i \in \beta_i, z_i \in \gamma_i)
\]

where \( 0 \leq i \leq n - 1 \), \( p_{i,m} \) is the scale ratio which is initialized as \( k_i = 1 \), while \( \alpha_i \), \( \beta_i \), and \( \gamma_i \) are the corresponding biomechanical constraints.

The extracted human figure from a monocular image is expressed as:

\[
X_{image} = (p_{0,i}(x_0, y_0), p_{1,i}(x_1, y_1), \\
\cdots, p_{n-1,i}(x_{n-1}, y_{n-1}))
\]

The projected 2D human figure from the 3D human model is represented as:

\[
X_{project} = (p_{0,p}(x_0, y_0), p_{1,p}(x_1, y_1), \\
\cdots, p_{n-1,p}(x_{n-1}, y_{n-1}))
\]

Suppose there are \( r \) joints that are directly connected with the \( i \)th joint of the human model, and they are \( p_{i1,m}, p_{i2,m}, \ldots, p_{ir,m} \). The scale ratio \( k_i \), which is used to adjust the lengths of the body parts that are linked with \( p_{i,m} \), can be computed as:

\[
k_i = \frac{\sum_{j=1}^{r} \left\| p_{i,j}(x_j, y_j) - p_{0,j}(x_j, y_j) \right\|}{\sum_{j=1}^{r} \left\| p_{i,j}(x_j, y_j) - p_{0,j}(x_j, y_j) \right\|}
\]

A monocular image containing human figure is used to demonstrate the aforementioned method, as illustrated in Fig. 2. Figure in upper left is the monocular image; figure in upper right is the extracted image features; figure in lower left is projected result of the human model in its initial posture; figure in lower right is the projection of the adjusted human model.

![Fig. 2. Adjustment of human model](image)
$EF_i = Scale(1) \cdot \Delta_{orientation_i} +$
$Scale(2) \cdot \Delta_{length_i} +$
$Scale(3) \cdot \Delta_{position_i},$

where $Scale$ is the matrix of weighting parameters.

Sometimes, adjustment of a joint can affect not only the segments directly linked to it, but also the orientations of other body parts. In view of this possibility, $EF$ of such a joint should take into consideration all its descendants. For example, assume that there are $t$ descendant joints of joint $i$, and they are joint $i+1$, joint $i+2$, ..., and joint $i+t$. $EF$ of joint $i$ should be:

$$EF_i = w_{i1} \cdot EF_{i1} + w_{i2} \cdot EF_{i2} +$$
$$\ldots + w_{it} \cdot EF_{it},$$

where $w_{i1}$, $w_{i2}$, ..., and $w_{it}$ are related weighting parameters for the descendant joints.

**IV. MOTION RECONSTRUCTION**

There are two types of movements applied on the 3D model in human motion reconstruction: translation and rotation. Translation determines the location of the human model, while the body posture is the result of rotations of every joint. Since joint Hip is the root of the connecting tree, translation of the human model can be implemented by translation of joint Hip.

**A. Translation of Human Model**

Joint Hip is translated for two times to get the global position of the human model, as shown in Fig. 4.

First, Hip is translated in the plane that contains joint Hip and is parallel to the image plane. The 1st translation of joint Hip places the projected point of Hip to the same position as the extracted point of Hip in the source image. $EF$ used in the 1st translation of joint Hip can be expressed as:

$$EF_{Hip} = \Delta_{position_{Hip}}$$

After the posture of the whole body is obtained by rotations of all the joints, Hip is translated for the second time along the line that is defined by two points: position of the camera and extracted feature point of joint Hip in the 2D image. The 2nd translation of joint Hip ensures that the projected human posture has the same size as the extracted human figure from the monocular image. $EF$ used in the 2nd translation of joint Hip can be expressed as:

$$EF_{Hip} = \sum_{j=0}^{i} (w_j \cdot \Delta_{position_j})$$

**B. Rotation of Joints**

Since the projected result of the initial posture of the 3D human model is often absolutely different with the extracted human figure, when one certain joint is being adjusted after the 1st translation of joint Hip, only the information of the segment(s) formed by connecting it with its immediate descendant(s) can be considered, and this method is named as Local Adjustment (LA).

For example, when joint Chest is dealt with, three segments, link between Chest and Neck, link between Chest and Luparm, and link between Chest and Luparm, are considered, as shown in Fig. 5. Since the distance between the human model and the camera is unknown yet, deviation of position cannot be considered when applying (5), i.e., $Scale(1)<0$, $Scale(2)<0$, $Scale(3)=0$. Thus $EF$ of joint Chest in LA is:

$$EF_{Chest} = w_{0/1} \cdot EF_{0/1} + w_{0/2} \cdot EF_{0/2} +$$
$$w_{0/3} \cdot EF_{0/3}$$

As adjustment of one joint can affect all its descendant joints, rotations should be applied to the joints from the root of the human connecting tree to the leaves, i.e., in such an order: Hip, Lthigh (and Rthigh), Lshin (and Rshin), Abdomen, Chest, Neck, Luparm (and Ruparm), Llowarm (and Rlowarm).

After LA, every joint can be rotated to its approximate relative position. Because the 2nd translation of joint Hip affects postures of the joints, they need to be rotated for another time. When one joint is being further adjusted, the information of all the segment(s) formed by connecting among it and all its descendant(s) should be considered, and this method is named as Global Adjustment (GA).

Obviously, those segment(s) directly linked to the joint are affected directly, while the other descendant(s) are indirectly affected. Thus weighting parameters of different joints should
be various, as illustrated in Fig. 1. Again take joint Chest as an example; as shown in Fig. 6, EF in GA is defined as:

\[
EF_{\text{Chest}} = c^2 * EF_{0/1} + c * EF_{1/2} + \\
c^2 * EF_{0/3} + c * EF_{3/4} + EF_{4/5} + \\
c^2 * EF_{0/6} + c * EF_{6/7} + EF_{7/8} 
\]

(10)

Since the distance between the camera and the human body is obtained after translation of human model, all the three factors of (5) should be considered in GA, i.e., Scale(1)<>0, Scale(2)<>0, Scale(3)<>0.

C. Reconstruction and Animation Procedure

In order to reconstruct the whole human posture from a single monocular image, a procedure is developed, during which translation of joint Hip and rotation of the joints are utilized, as shown in Fig. 7. If the accuracy of the reconstructed human posture is still not satisfied, more times of GA can be executed.

Based on physiology knowledge, due to the passive viscoelastic and active chemico-mechanical properties of human muscles, the muscle movement is constrained to proceed in a relatively smooth and continuous fashion where abrupt changes are not permitted. Using this rule, the 3D postures reconstructed from all single frames can be combined to constitute a legal motion.

Obviously, 3D posture of the human body is mainly determined by the rotational angles of all joints. Hence smooth and continuous changes between two neighbor postures can be represented by small changes of the joint angles. In order to define EF for a series of monocular images, two more notations are used:

\[
V_{i,j-1..m} \text{ is the vector of the } i\text{th segment of the 3D human posture reconstructed from the } (j-1)\text{th frame, and} \\
V_{i,j..m} \text{ is the corresponding vector of the } i\text{th segment of the reconstructed human posture from the } j\text{th frame.}
\]

Therefore, the change between these two vectors can be expressed as:

\[
\Delta_{\text{neighbor}} = \max \left( \frac{V_{i,j-1..m} \cdot V_{i,j..m}}{\left| V_{i,j-1..m} \right| \left| V_{i,j..m} \right|} \right) 
\]

(11)

The EF used for image sequences can be defined as:

\[
EF_i = \text{Scale}(1) * \Delta_{\text{orientation}}_i + \\
\text{Scale}(2) * \Delta_{\text{length}}i + \\
\text{Scale}(3) * \Delta_{\text{position}}_i + \\
\text{Scale}(4) * \Delta_{\text{neighbor}}_i 
\]

(12)

If the current image is the first one of the monocular image series, there should be \( \Delta_{\text{neighbor}}_1 = 0 \).

To animate a series of human motion from a monocular video sequence, a model-based human animation system is developed based on the techniques discussed above. In the animation procedure, (12) is used as the EF to ensure the smooth transition of postures between the neighbor frames.

1. Take the feature points extracted from a series of monocular images as input.
2. Adjust relative lengths of the segments according to the obtained 2D information.
3. Calculate the initial projection value of the human model for every body part.
4. Determine 3D location and orientation of every joint and segment of the human model using the reconstruction procedure illustrated in Fig. 7, during which (12) is applied as EF instead of (5).
5. Repeat Step 3 and 4 for all the other frames.
6. Show the reconstructed 3D postures consecutively.
7. End.

V. EXPERIMENTAL RESULTS

Our posture reconstruction procedure is tested using some monocular images downloaded from the Internet. As shown in Fig. 8, figures of the first row show the source monocular images with manually picked feature points, while the second row displays the reconstructed postures from the same viewpoint. The third row shows the results from a 45° rotated viewpoint, and the fourth row shows the results from a 90° rotated viewpoint.

Next the animation procedure is tested by a monocular video sequence of a dancing person. Feature points in the first frame are picked manually, while image processing techniques such as Linear Prediction, Normalized Cross Correlation, Least Square Matching etc. are used to help extract 2D information of the features from the other frames. As shown in Fig. 9, figures in the 1st column are monocular frames of the video sequence;
figures in the 2nd column are the extracted feature points (dot) and the animated results (solid) from the same viewpoint; figures in the 3rd and 4th column are results from side view and top view respectively. There are 50 frames in this dancing series, and 8 of them (10th, 15th, 20th, 25th, 30th, 35th, 40th, 45th) are selected and displayed.

VI. CONCLUSION

A model-based approach is presented in this paper for motion recovery from un-calibrated monocular images that contain unrestricted human postures and movements. In order to obtain the minimum value of the Energy Function, translations and rotations (Local Adjustment and Global Adjustment) are proposed to adjust the flexible 3D human model with encoded biomechanical constraints.

The advantage of this approach is that it can deal with a wide range of monocular images, e.g., obtained from Internet, scanned photographs, historical shots, or clips of films, and neither camera calibration nor user interface is needed during the reconstruction procedure.

Experiments show that the results are well acceptable, while improvements are still needed, such as automatic picking of occluded feature points, adjusting unnatural reconstructed postures, further studying of biomechanical constraints and Energy Function, etc.

REFERENCES

Fig. 9. Animation of a dancing sequence