Video Motion Capture Using Feature Tracking and Skeleton Reconstruction

Yueting Zhuang, Xiaoming Liu and Yunhe Pan
Institute of Artificial Intelligence
Zhejiang University
Hangzhou 310027, P.R.China
{yzhuang,liuxm}@icad.zju.edu.cn and panyh@sun.zju.edu.cn

Abstract

In the domain of computer vision, there exists a very wide application for the research of human motion capture. This paper proposes a new approach to do motion capture in video. It is composed of image sequence based tracking of human feature points and the reconstruction of three-dimension(3D) motion skeleton. First, we track every part of human body from top to bottom on the basis of a human model. The Kalman Filter and a morph-block similarity algorithm based on subpixel are used. Then we do camera calibration using the line correspondences between the 3D model and the image. Finally the 3D motion skeleton is established by use of the model knowledge. This approach does not aim at a given mode of human motion. Rather, it analyzes large motion from frame to frame in complex, variational background, and sets up a 3D motion skeleton under the perspective projection. We also present the experiment result at the end of the paper.

1. Introduction

Motion capture plays an important role in the creation of special effects in many fields. Except for the film and animation, motion capture has a comprehensive application in the analysis of athlete performance, medical diagnostics, surveillance, and video retrieval etc.

The conventional motion capture has two approaches. One is to attach many sensors to the joints of human body. It will record the position of joint at every time. The other is to analyze a video in the following three steps: 1) feature extraction in video frames, 2) correspondence between the features of every frame, 3) recovery of 3D motion from feature correspondences. O'Rouke and Badler[11] analyze 3D human motion by mapping the input images to a volumetric model. In the systems of Hogg[6] and Rohr[10], edge and line features are extracted from images and matched to a cylindrical 3D body model. Chen and Lee[5] use 17 line segments and 14 joints to represent the human skeleton model. Bharatkumar et al.[3] also use stick figures to model the lower limbs of the human body. Their goal is to construct a general model for gait analysis in walking. Bregler and Malik[4] recover the information of 3D human motion under the orthographic projection by marking limb segments in the initial frame. For the special complexity of human motion, the existing research methods lay much limitation on human[1], such as a single and quiescent background, parallelism of motion direction to the image plane, and tight clothing of human. To attach sensors will cost too much money and time, restrict free movement.

We propose a new approach to do motion capture. Our approach removes many restrictions as in the previous approaches. It does not aim at a given human motion mode. Rather, it analyzes large motion from frame to frame in complex, variational background, and finally sets up a 3D human skeleton under the perspective projection. Then this model can be used in many applications such as human animation, VR, etc. The only need for the user is to mark the joints of the first frame and computer will do the rest. In particular, we emphasize on two points. One is to acquire the sequence of 2D human motion skeleton by tracking joint with the support of motion prediction and color model of body part. The other is to use the correspondences between the 3D model and the 2D image to calibrate camera and establish the sequence of 3D human skeleton under the perspective projection.

The architecture of video motion capture is shown in figure 1. The content in the dashed boxes is the two focuses mentioned above. This paper is organized as follows. Section 2 introduces the human model used in our approach. The human skeleton tracking of image sequence and reconstruction of 3D human motion skeleton sequence are detailed in section 3 and 4 respectively. Section 5 shows the experiment result. Finally we give the conclusions.

2. The Human Model

The basic idea is to regard the 3D human body as an articulated object[7] and simplify the human motion to the

motion of skeleton. Figure 2.a shows a 3D human skeleton model. It contains 16 joints, which are named 3D feature points. From the knowledge of anatomy, we can acquire the length proportion of each line in this model.

This paper names the projection of a 3D feature point as a 2D feature point. In the tracking of 2D image sequence, we use block to represent the projection of a body part in the image plane (see figure 2.b). The middle line of each block is the skeleton after projection. It divides a corresponding block into two small blocks of equal area. After the marking of the first frame, we may get color model of each block. By searching out its new position in the subsequent frames, we could track the human skeleton on the image plane.

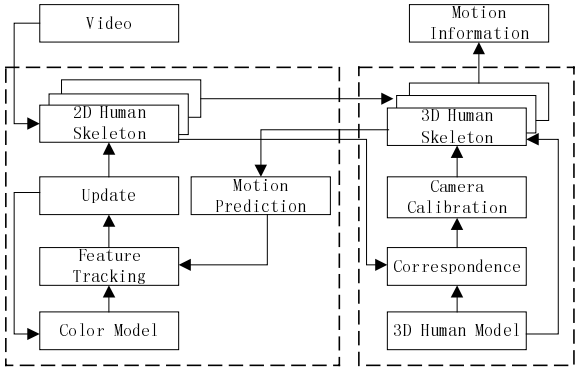


Fig.1. The architecture of video motion capture

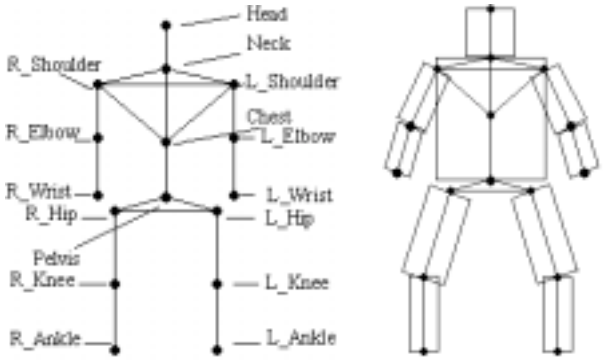


Fig.2. Human model: (a) 3D skeleton model(left) (b) 2D block model(right)

3. Skeleton Tracking

Because there is little self occlusion on human head, we can acquire its color information easily. So, beginning with head, we track every body part from top to bottom. Now we detail the tracking of head, trunk and limb respectively.

3.1. Head

For every frame in the sequence, the head may move toward any directions in the next frame. To reduce the search area of calvaria point in the next frame, we introduce Kalman filter based global motion model to predict the calvaria point. Then we select a search path to do morph-block match around the predicted point.

Kalman Filter. Regarding the sequence of motion images as a dynamic system[9], the calvaria point can be described by the following equation: $P=P' + \eta$. The coordinate $P=(x,y)^T$ is the tracked calvaria point. P' is the actual coordinate. η is a 2D gaussian random noise with mean value 0 and covariance R. We use thrice polynomial to represent the motion trajectory of point P. The state vector is defined as $S=(P,P',P'')$, where $P'=(x',y')^T$,and x',y' represent the velocity of point P in the X,Y directions respectively. $P''=(x'',y'')^T$, where x'',y'' represent the acceleration of point P in the X,Y directions respectively. The state equation is defined as

$$S(k+1) = F \quad S(k) + G \quad n(k) \tag{1}$$

where $F = \begin{bmatrix} I_2 & I_2 \bullet T & \frac{1}{2} I_2 \bullet T^2 \\ 0_2 & I_2 & I_2 \bullet T \\ 0_2 & 0_2 & I_2 \end{bmatrix}$ $G = \begin{bmatrix} \frac{1}{2} I_2 \bullet T^2 \\ I_2 \bullet T \\ I_2 \end{bmatrix}$

$K=0,1,2,\ldots$ represents the serial number of the frame, I_2 is a 2×2 unit matrix, 0_2 is a 2×2 zero matrix, and T is time interval between frames. $n(k)=(n_x(k),n_y(k))^T$ describes the acceleration noise in the x,y directions. We suppose $n(k)$ conforms to gaussian distribution with even 0 and covariance Q. This state equation shows that P is doing varied-acceleration linear motion in all the X,Y directions. In practice, we track the coordinate of point P, namely $X(k)=p(k)$. So the measurement equation is:

$$X(k) = H \bullet S(k) + \eta(k) \tag{2}$$

where $H=[I_2,0_2,0_2]$ is a 2×6 matrix. In the above condition, we get the recursive equations of kalman filter[8]. Kalman filter consists of initialization, prediction and update. The flow chat is shown in figure 3. Our experiments show that using kalman filter to predict the calvaria point has a good performance.

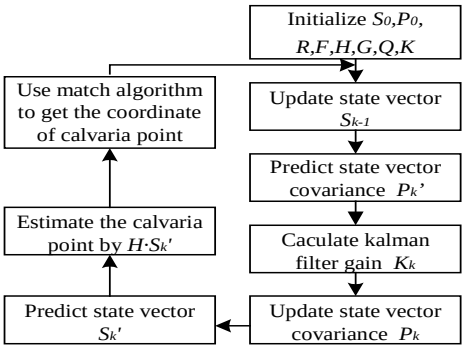
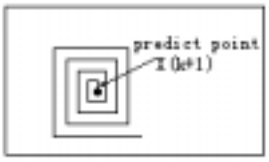


Fig.3. The flow chart of Kalman filter

Morph-block match. We have applied the kalman filter to predict the calvaria point in the next frame. Now we will choose a search path (figure 4) to do morph-block based match around the predicted point.



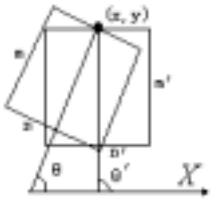


Fig. 4. Search path Fig. 5. Two feature morph-blocks

Because we have known the calvaria and neck point in the first frame, the height *(m)* of head block is the distance between these two points and the proportion of height to width*(n)* can be acquired in anatomy. The color information of $m \times n$ pixels in this block is saved as the color model for the match of subsequent frames. Since the head block in the image is the projection of human head, the head motion will change the shape of projection. For example, the head block becomes larger which is likely to happen when a human is moving toward the camera. So, the block match must be processed between morph-blocks. For this, we propose a *morph-block weighted similarity algorithm based on subpixel*.

Define a feature morph-block $A=\{(x,y),m,n,\theta\}$(see figure 5), where *(x,y)* is the intersection of one side and the middle line, *m* is the height of block *A*, *n* is the width of block, and θ is the angle between the middle line and *X* axis. Now there are a reference block $A=\{(x,y),m,n,\theta\}$ and a comparative block $A'=\{(x',y'),m',n',\theta'\}$. To calculate their similarity, use the algorithm as below:

1. If $m \times n < m' \times n'$, Then *row=m*, *column=n*; Else $row=m'$, $column=n'$;
2. In block *A* we depict *column* and *row* pieces of gridding lines evenly in the direction of $arctg\ \theta$ and $arctg(-1/\theta)$ respectively. We name the intersection of any two gridding lines as subpixel X_{ij},$(0 \leqslant i < m, 0 \leqslant j < n)$. Then we use quadric linear interpolation to calculate the color of every subpixel, $X_{ij}[Red]$, $X_{ij}[Green]$, $X_{ij}[Blue]$.
3. In block A' we depict *column* and *row* pieces of gridding lines evenly in the direction of $arctg\ \theta'$ and $arctg(-1/\theta')$ respectively. Then we use linear interpolation to calculate the color of every subpixel, $X_{ij}'[Red]$, $X_{ij}'[Green]$, $X_{ij}'[Blue]$.
4. Calculate:

$$\begin{aligned} diff_{ij} &= W_R \bullet |X_{ij}[Red]-X_{ij}'[Red]| \\ &+ W_G \bullet |X_{ij}[Green]-X_{ij}'[Green]| \\ &+ W_B \bullet |X_{ij}[Blue]-X_{ij}'[Blue]| \end{aligned} \quad (3)$$

$$S = 1/(W_1 \bullet \sum_{(i,j)\in b1} diff_{ij} + W_2 \bullet \sum_{(i,j)\in b2} diff_{ij}) \quad (4)$$

where W_R, W_G, W_B represent the weight of each element in *RGB*, *b1*,*b2* represent the two regions divided in the block, W_1, W_2 represent the weight of each region in the whole block. In the case of the head, we define the center region as *b1* and the margin region as *b2*, namely:

$$\begin{cases} (i,j)\in b1 & If \quad m/4 \le i \le (3/4)m \quad AND \quad n/4 \le j \le (3/4)n \\ (i,j)\in b2 & Otherwise \end{cases} \quad (5)$$

Here we have $W_1 > W_2$. This weighted morph-block similarity measure is based on the observation that the margin region of head has a more salient change of color in motion, however the center region has a relative small change. *S* is used to represent the similarity of two morph-blocks. Apparently, the bigger *S* is, the larger the similarity is.

For a frame sequence, we define the tracked head block in current frame as the reference block *A*, and the head block in the next frame as the comparative block A'. We set θ' as $\theta - \Delta\theta \leqslant \theta' \leqslant \theta + \Delta\theta$ and m' as $m - \Delta m \leqslant m' \leqslant m + \Delta m$. Since the height and width of head zoom in proportion, we set n' as $n-(n/m)\Delta m \leqslant n' \leqslant n+(n/m)\Delta m$. Then for every point *(x,y)* on the search path, we form several A' by $\{(x,y),m',n',\theta'\}$ and calculate its similarity with head block of current frame, *A*. The system records the A' which has the largest similarity. After finding the largest similarity on the past search path, the search process will continue until on the search path of next one circle it does not find a point which has a larger similarity. If does, repeat the process mentioned in the last sentence. In the end, the recorded A' is the head block of next frame. And for the self adaptability of color model, we utilize linear weight to update the color model[2].

3.2. Trunk and Limb

When the head block is tracked, one feature point of trunk, the neck, is fixed on. The tracking of trunk and limb also depends on the above algorithm. But we must pay attention to two problems. Firstly, because of the large limb motion from frame to frame, we introduce a prediction mechanism to estimate the potential limb position in the next frame, and then fix it on accurately[8]. Secondly, we show how to deal with self occlusion in the tracking of limb. For example, there is relative small similarity in the block match when in one frame the trunk occludes an up limb. But the similarity will be larger as soon as the occlusion disappears. According to this, we also define the similarity *S* as the reliability of block match. In the match process of frame sequence, we preserve the reliability of every limb match. If there are one or several low reliability frames between two relative

high ones, we use the joint coordinate of high ones to obtain the joint of low ones by linear interpolation. Our experiment shows that it can deal with self occlusion to a certain extent and optimize the tracking performance.

4. Reconstruction of 3D human motion skeleton

To establish the sequence of 3D human motion skeleton under the perspective projection, we must first acquire the camera parameter, namely camera calibration in computer vision. Then we calculate the coordinate of 3D feature points on the human model by use of the pin-hole model and the knowledge of human skeleton.

Consider two coordinate systems[9], $O_wX_wY_wZ_w$ and $O_cX_cY_cZ_c$. The former is an object space coordinate system in which the 3D feature points are located. The camera is referenced as the camera coordinate system $O_cX_cY_cZ_c$. Then every point P_w in $O_wX_wY_wZ_w$ can be translated to (u,v) on the image plane by two transformations, a rotation R and a translation t. Our goal in camera calibration is to determine R and t when some corresponding feature lines between 3D human model and image plane are given. To simplify the calculation of partial derivative, we transform the standard formulas into

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = R \begin{pmatrix} X_w \\ Y_w \\ Z_w \end{pmatrix} \tag{6}$$

$$(u, v) = \left(\frac{f \bullet X'}{Z' + Dz} + Dx \, , \frac{f \bullet Y'}{Z' + Dz} + Dy \right) \tag{7}$$

We substitute translation t with Dx,Dy,Dz, and represent parameter R by a rotation vector, $(Wx,Wy,Wz)^T$. We define the equation of a line, with a point (u,v) on it, by

$$\frac{-m}{\sqrt{m^2+1}} u + \frac{1}{\sqrt{m^2+1}} v = d \tag{8}$$

where d is the perpendicular distance from the origin to that line, and m is the line slope. Now the partial derivative of d to Dx,Dy,Dz,Wx,Wy,Wz will be obtained. After that, we may use Newton method to obtain six projective parameters based on the following equation:

$$\frac{\partial d}{\partial Dx}\Delta Dx + \frac{\partial d}{\partial Dy}\Delta Dy + \frac{\partial d}{\partial Dz}\Delta Dz + \frac{\partial d}{\partial Wx}\Delta Wx + \frac{\partial d}{\partial Wy}\Delta Wy + \frac{\partial d}{\partial Wz}\Delta Wz = E \tag{9}$$

where E is the perpendicular distance from the end points of a 2D feature line to the projective line. Because there are two end points on one line, we can get two equations such as (9) for one pair of corresponding feature lines. Given three pairs of such lines, six equations will form a linear equation group. So, there are at least three pairs of corresponding lines needed in Newton method. In the human model, we choose the line between left and right shoulders, and the two lines between the chest and two shoulders. This choice is based on the observation that this triangle should not morph-itself in motion under most conditions. In the below description, we name each feature object of this triangle as the key joint, key line, and key triangle. In the first frame, the projection of key joints on the image plane is known by manual marking. The position of key joint in the object space coordinate is specified by our system. As long as the proportion of each key line accords with the anatomy, we can always find the location and orientation of the camera in the object space coordinate system and let the perspective projection of key triangle superpose with the up triangle of trunk on the image plane.

Now corresponding to the first frame, except for three key joints, all other 3D feature points of the human model are not determined yet. The next step is to acquire the 3D feature point coordinate (X_c,Y_c,Z_c) of the human model corresponding to a known 2D feature point coordinate, (u,v). As known from the pin-hole model, to link the optical center and a projective point will get a line, on which all the points project on the same point in the image plane. In order to locate the 3D feature point on this line, we begin with a known neighboring point and use the knowledge of human skeleton length to find a point, the distance from which to the known neighboring point is equal to the corresponding skeleton length[8]. Thus, with the order from center to fringe in skeleton model, we can get all the 3D feature points coordinates in the human model.

Then we discuss how to determine the coordinates of three key joints corresponding to the subsequent frames[7]. Given the key joint coordinates, $P_i^n(X_i^n,Y_i^n,Z_i^n)$ $(i=1\sim3)$, of frame n in the camera coordinate system, let us calculate the corresponding key joint, P_i^{n+1} $(X_i^{n+1},Y_i^{n+1},Z_i^{n+1})(i=1\sim3)$, of frame $n+1$. The corresponding 2D feature point in the image plane is (U_i^{n+1},V_i^{n+1}). The relation of P_i^{n+1} and (U_i^{n+1},V_i^{n+1}) can be described as

$$P_i^{n+1} = \left(\frac{U_i^{n+1} \bullet Z_i^{n+1}}{f}, \frac{V_i^{n+1} \bullet Z_i^{n+1}}{f}, Z_i^{n+1} \right) (i=1\sim3) \tag{10}$$

The skeleton length in the human model is invariable, which means:

$$d(P_i^n, P_j^n) = d(P_i^{n+1}, P_j^{n+1}) \; (i,j=1\sim3) \; And \; i \neq j \tag{11}$$

Using (10) to substitute P_i^{n+1} in (11), we will get a nonlinear equation group, which has three variables and may be solved by the grads method. Thus, we obtained the key joint coordinates of frame $n+1$ in the camera coordinate system.

Finally we can calculate all the 3D feature points in human model corresponding to frame $n+1$ by the algorithm used on the first frame[8].

5. Experiment results

We have implemented a demo system on personal computer. Our experiment materials are some video

recordings shot with a single camera. Figure 6 shows an example of a human walking toward the camera. In the top row, the 16 feature points on the first image are marked by the user with the mouse. After the hand-initialization we applied the program to a sequence of 41 image frames. We could successfully track all body joints in the frame sequence. The other four images of the top row are the 5th, 17th, 23rd and 41st frames of the video respectively. In the middle row, we show five images of our constructed 3D human skeleton which is shot with the same view point as the top ones. In the bottom row, these five images are shot by the camera rotating left 15°, right 45°, right 30°, left 60° and left 30° respectively. As you see, the motion continuity and authenticity are embodied in the walking sequence, which proves the robustic algorithm of 3D reconstruction. It means now we can see the person walk from every viewpoint. For more information of our experiment results, please visit: http://icad.zju.edu.cn/ ~liuxm /animation.html.

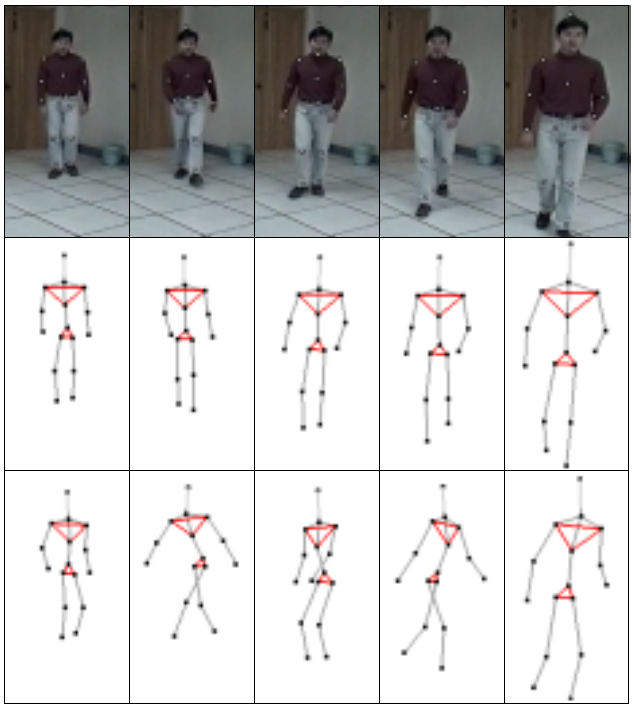


Fig.6. The experiment of human walking

6. Conclusions

This paper proposes a new technique to capture motion in video. It is a challenging domain to track human motion in the joint level and recover 3D motion information. Our contribution to this problem is that this approach does not pose any restrictions on human motion and finally set up a 3D motion model under the perspective projection. To the best of our knowledge, this is the first demo system that is able to process such a challenging task and recover complex human motion with high accuracy. It is easy and straightforward from a user's point of view. The only need for the user is to mark the joints of the first frame and computer does the rest. On the other hand, any video stream, whether it is a film, or any historical shot, such as Charlie Chaplin's walking and Karl Lewis' running, can be our material, which means the comprehensive application foreground of our approach. Future work will concentrate on utilizing more 3D human skeleton motion knowledge to guide the 2D feature tracking. We should also implement the texture mapping to visualize the human skeleton.

Acknowledgements

Our work was sponsored by the National Natural Science Foundation of China. We would also like to thank Yiying Tong, Kun Zhou, Yi Wu and Xiqun Lu for fruitful discussions.

References

[1]. J. K. Aggarwal, and Q. Cai, “Human Motion Analysis:A Review”, In Proc. of the IEEE Nonrigid and Articulated Motion Workshop, Piscataway, NJ, 1997, pp.90-102.

[2]. K. Akita, “Image Sequence Analysis of Real World Human Motion”, Pattern Recognition, Vol.17 No.1,1984, pp.73-83.

[3]. Bharatkumar, A. G., Daigle, K. E., Pandy, M. G. , Cai, Q. and Aggarwal, J. K., “Low limb kinematics of human walking with the medial axis transformation”, In Proc. Of IEEE Workshop on Motion of No-Rigid and Articulated Objects, Austin,TX,1994, pp.70-76.

[4]. C. Bregler and J.Malik, “Video Motion Capture”, In Proceeding of SIGGRAPH 98.

[5]. Z.Chen and H.Lee,“Knowledge-guided visual perception of 3D human gait from a single image sequence”, IEEE Trans. On Sys. Man. and Cybernetics, 22(2),1992,pp.336-342.

[6]. D.Hogg, “A program to see a walking person”, Image Vision Computing, 5(20),1983,pp.28-42.

[7]. T. S. Huang and A. N. Netravali, “Motion and Structure From Feature Correspondences: A Review”, Proceedings of The IEEE, Vol.82 No.2, February 1994, pp.252-268.

[8]. Liu XiaoMing, Zhuang YueTing, Pan YunHe, and Yang Jun, “Human Three Dimension Motion Skeleton Reconstruction of Image Sequence”, Accepted for publication in Journal of Computer Aided Design and Computer Graphics. Spring 2000.

[9]. Ma SongDe, and Zhang ZhengYou, Computer Vision: Compute Theory and Arithmetic Foundation, Scientific Press, Beijing, January 1998.

[10].K. Rohr, “Incremental recogniton of pedestrians from image sequences”, In Proc. IEEE Conf. Comput. Vision and Pattern Recogn., New York, June,1993, pp.8-13.

[11].J.O'Rourke and N.I.Badler, “Model-based image analysis of human motion using constraint propagation”, IEEE PAMI,2(6),Nov.1980,pp.522-536.