Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp458-463)

Human Animation from 2D Correspondence Based on Motion Trend Prediction

LI ZHANG
Department of Computing
Curtin University of Technology
Perth, WA 6845

Australia

LING LI
Department of Computing
Curtin University of Technology
Perth, WA 6845

Australia

Abstract: A model-based method is proposed in this paper for 3-dimensional human motion recovery, taking un-calibrated monocular data as input. Motion trend prediction is suggested to to recover smooth human motions with high efficiency; while its outputs are guaranteed to not only resemble the original motion from the same viewpoint the sequence was taken, but also look natural and reasonable from any given viewpoint. To evaluate the accuracy of reconstruction algorithm, the research program starts from "synthesized" input. Experiment carried on real video data will be discussed as well, which indicate that the proposed method is able to recover smooth human motions from their 2D image features with high accuracy.

Keywords: smooth human animation, continuous rotation, 2D monocular data, motion trend prediction

1. Introduction

Animation is the production of consecutive images, which, when displayed, conveys a feeling of motion [9]. In the past decade, with the rapid development of computer technology, computer animation has become very popular in many applications. In computer animation, the representation of human body and its motion receives great attention, since human animation are widely employed in many areas, such as games, movies, surveillance, scientific visualization, etc. As monocular images and video sequences are easily available, many great efforts have been made to reconstruct 3D human motion from monocular images. However, such attempt remains very much under-developed due to many technical difficulties.

Remondino and Roditakis [14] suggested adjusting the posture of a human model according to camera calibration information and biomechanical constraints applied on the model. Orthographic projection is used in their approaches, which is greatly different from the perspective projection used in any real camera. Liu et al. [12] and Park et al. [13] made use of the motion library. The former took motion attributes achieved through reconstruction as guidance for estimation of unknown human motion, while the latter use motion library to resolve the depth ambiguity in recovering 3D configuration from 2D image features. In both attempts, a large motion library needs to be maintained and upgraded continuously. In [3] the concept of prioritized constraints is introduced. Based on it the proposed method can get quite good results, which is only suitable for adding variations to motions known before the reconstruction. [4] introduced an interactive system which combines biomechanical constraints on 3D motion with user interferences to reconstruct sequences in 3D; similarly three possibilities for solving inverse kinematics problem during human animation are discussed when interactive direct manipulation is applied [5]. There are also attempts in automatically generating accurate inverse dynamics solutions to simulate and deform human motion [11&16]; however such efforts have been concentrated on hand posture recovery only. Zhao and Li [18] proposed a Criterion Function (CF) to represent the residuals between the image feature and the projected features from the reconstructed 3D model in a Global Adjustment (GA) system. In their method the accuracy and the consistency of the recovered postures are only guaranteed from the same viewing direction as the original.

Most existing methods introduce simplifications on human motion or require assistance such as user interference or motion library. This paper aims to propose a novel model-based human motion reconstruction method from un-calibrated 2D monocular data without user interferences and the human motion is truly unrestricted.

The rest of this paper is organized as follows: the camera model and the 3D skeletal model used for motion reconstruction are described in the next section; section 3 discussed the key component of our work: the MTP method; experimental results are presented and analyzed in section 4; finally a conclusion is drawn in section 5.

2. Model

Before 3D motion reconstruction, we have a camera model and a human model well prepared for the purpose. The camera model is located at a fixed position in virtual space with pre-defined focal length, and it does not require knowledge of the actual cameras from which the video sequence is taken (in most of the situations such information is unavailable as well). An articulated 3D skeletal model as shown in Fig.1 is considered sufficient for our purpose. For convenience terms in italic stands for joints, while segments' name are underlined. The joint *pelvis* is set as the root in the skeletal tree structure, while the 5 leaf joints are named *left wrist*, *right wrist*, *left ankle*, *right ankle* and *head*.

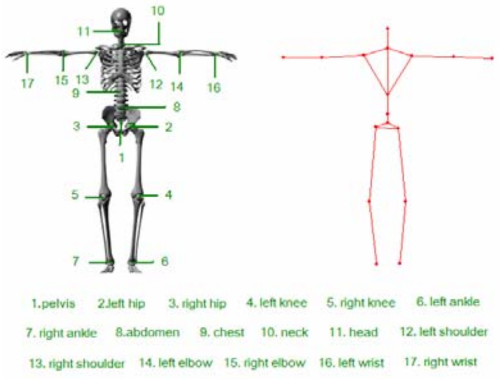

Figure 1: 3D skeletal model and its 2D correspondence composed of 17 joints and 12 segments

At the current moment, our attention is focused on the movements of segments including hip, waist, chest, neck, upper arms, forearms, thighs, and crura, which are resulted from proper rotations about each starting joint of these segments. The tips of hands and feet, and the top of the head are not considered as individual joints in our current work. For reconstruction purpose, we assign each intermediate joint (all joints in Fig.1 except leaf joints) 1 to 3 DOF(s) in the world coordinate system (WCS), and each of these joint is associated with a local coordinate system (LCS). However there are 6 DOFs (3 for translations and the other 3 for rotations) at *pelvis*, since the translation of the whole body is represented by the movement of hip. In total our human model actually 37 DOFs.

It's obvious that the rotational angles of each intermediate joint about the 3 axes in its associated LCS are governed by biomechanical constraints [17], here we decide to introduce the joint constraints in [18] with some modifications applied according to latest human biology information.

Through proper translations of the *pelvis* joint and rotations of segments about each intermediate joint of the skeletal model, any human posture could be obtained from the initial pose in Fig.1.

3. Motion Reconstruction

Human motion reconstruction is actually a process to reconstruct human posture at every frame. The goal is to recover human postures which resemble the original postures as much as possible. Our algorithm is based on two observations about human motion. 1. Despite the complexity of human motion, there are actually only two types of movements involved: the translation of the whole body and the rotations of the body segments about each intermediate joint. The former put the human body in a certain location, while the latter generates a particular body posture. Hence, the 3D movement of any joint can be treated as the composition of translation of *pelvis*, and rotations of all ancestor joints of this particular joint. 2. Most of the human motions are generally smooth, which means that the 3D human postures in neighboring frames are similar. Based on the above observations, the Motion Trend Predication technique is proposed.

3.1. The MTP technique

In the early research stage, only monocular human motions without body relocation are studied. Therefore the joint *pelvis*'s position is always fixed and the movement of each joint can be simplified just as a composition of the rotations about all its ancestor joints.

According to kinematics principles [10], we will follow the order from the root to the leaf joints to adjust segments of the skeletal model. As mentioned before, once human body gets located a particular posture could be obtained through proper rotations of each segment about its starting intermediate joint. Therefore only when the 2D residuals between the recovered posture's projection and its corresponding 2D image feature achieve the minimum value, can we consider all joints and segments have been transformed to certain 3D positions which merely ensure the recovered posture resembles the original one from the same viewpoint. However recovering posture at each frame individually may violate the inter-frame consistency. To effectively produce natural looking motions, an important 3D predictor is introduced in the suggested MTP. Since the 3D positions of any joint in neighboring frames should be similar to ensure continuous motion, coordinates of every joint in the WCS could be utilized for motion prediction and tracking. When a segment is being rotated about the i^{th} intermediate joint, WCS coordinates of its all child joints will serve as the 3D predictor to better evaluate if the segment has been rotated to the correct location. To obtain the value of this 3D predictor, below formula is applied, which represents track of a certain joint's movement with frames:

$$predictor_{3D\_pos\_i} = \sqrt{(x^K - x^{K-1})^2 + (y^K - y^{K-1})^2 + (z^K - z^{K-1})^2}$$

$$\left(K \geq 2\right) \qquad (1)$$

where (x^K, y^K, z^K) and (x^{K-1}, y^{K-1}, z^{K-1}) stand for each i^{th} joint's direct descendant's WCS coordinates at the K^{th} and $(K+1)^{th}$ frame respectively.

Once the reconstruction of a frame is finished, the posture configuration of the skeletal model obtained for this frame will be used as the reference for predicting the human posture in the next frame.

To represent the MTP in a parametric way, an AF for optimizing rotation about the i^{th} intermediate joint is declared as shown in Eq.(2).

$$\begin{aligned} AF_i = {} & weighting\_parameter_{orientation\_i} \times deviation_{orientation\_i} \\ & + weighting\_parameter_{position\_i} \times deviation_{position\_i} \\ & + weighting\_parameter_{length\_i} \times deviation_{length\_i} \\ & + weighting\_parameter_{3D\_pos\_i} \times predictor_{3D\_pos\_i} \end{aligned} \tag{2}$$

where $deviation_{orientation\_i}$ and $deviation_{length\_i}$ represent 2D orientation and length deviations between projected and image features of the segment(s) connecting the i^{th} intermediate joint and its direct child joint(s), while $deviation_{position\_i}$ is 2D position deviation of the i^{th} intermediate joint's all direct child joint(s) on image plane.

Such above AF looks similar to the popularly-used energy function defined in [17]; however usage of 3D predictor for motion tracking and its trend prediction has been added in. This improvement is to enhance the accuracy of human animation in 3D space, which is the main concern in any 3D motion reconstruction. Meanwhile the way the AF is formulated in our MTP makes the solution process much simpler than all currently existing methods.

However, the current AF is still insufficient, especially when dealing with limbs, as possible ambiguities could be resulted from occlusion due to the high flexibility of limbs. To smooth such ambiguities, we assume that poses of the forearms or shins remain exactly the same as in the previous frame when <u>upper arms</u> or <u>thighs</u> are being rotated in current frame. The purpose is to combine the 2D position residuals of the *wrists* or *ankles* between the projection and image features into current AF. Hence the AF for rotating <u>upper arms</u> or <u>thighs</u> evolves as follows:

$$\begin{aligned} AF_i^{\lim b} = {} & AF_i \\ & + weighting\_parameter_{leaf\_jo\mathrm{int}\_i} \times deviation_{leaf\_jo\mathrm{int}\_i} \end{aligned} \tag{3}$$

$$deviation_{leaf\_jo\mathrm{int}\_i\_} = D_2\left(P_{j\_i}, P_{j\_p}\right) \tag{4}$$

Here the j^{th} joint (leaf joint – *wrist* or *ankle*) is an indirect descendant of the i^{th} intermediate joint (*shoulder* or *hip*); P_{j_i} is the image feature of the j^{th} joint, and P_{j_p} is its corresponding projected feature, as illustrated in Fig.2.

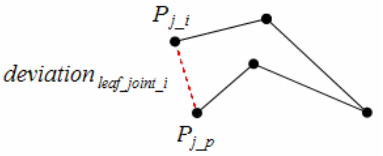

Figure 2: 2D position deviation of *shoulder* or *hip*'s indirect child joint (leaf joint)

Once the posture configuration resulting in the minimum AF values for every intermediate joint is obtained, all joint are believed to have been moved to their approximately correct 3D positions. As the result, the whole reconstructed human motion will be smooth, and looks natural from any viewpoint in 3D space.

Next step we attempt to handle unrestricted human motions containing body relocation. That means the relocation of the human body which is represented by the translations of the *pelvis* has to be taken into consideration as well. Such relocation can be further categorized into translations parallel and perpendicular to the image plane. The former is easy to implement, since in this case the distance between the human object and the camera is fixed. The AF for recovering *pelvis*'s parallel translations is defined as below:

$$AF_{translation\_1} = deviation_{position\_pelvis} \tag{5}$$

$$deviation_{position\_pelvis} = D_2\left(P_{pelvis\_i}, P_{pelvis\_p}\right) \tag{6}$$

where P_{pelvis_i} and P_{pelvis_p} are image feature and projection of *pelvis* respectively.

In contrast, the second type of translation is much more difficult to handle since it is not easy to detect when and how such translation exactly happens. It is a common knowledge that the distance of an object with the projection plane determines its projected sizes on the image plane. Such understanding is used to determine the translation of the *pelvis* perpendicular to the image plane. Before the *pelvis* being translated in the $(K+1)^{th}$ frame, the 3D posture configuration obtained at the K^{th} frame will be applied to the skeletal model temporarily. If the skeletal model is translated to its correct location at the $(K+1)^{th}$ frame, the sum of the projected segment lengths should approach to that of the image features as much as possible. The following AF is then defined to perform the body translation perpendicular to the image plane:

$$AF_{translation\_2} = deviation_{total\_length} \tag{7}$$

$$deviation_{total\_length} = abs\left(\sum_{i=1}^{16}\left\|\overrightarrow{S_{i\_i}}\right\| - \sum_{i=1}^{16}\left\|\overrightarrow{S_{i\_p}}\right\|\right) \tag{8}$$

where S_{i_i} and S_{i_p} respectively represent lengths of the image feature and projected feature of each segment considered in our model.

After the translation of the joint *pelvis*, the 3D skeletal model is considered as positioned to the right location in 3D space. Rotations of every intermediate joint will then follow using the methods described upfront. As the *pelvis*'s position changes in most of the frames, to recover the human posture through the same AFs for rotations about each intermediate joint, a new reference coordinate system (RCS) is introduced and serves as the substitute of the "WCS" for calculating the 3D predictor's value in MTP. Such RCS is defined after the translation of *pelvis* in every frame, with its origin located at *pelvis* and its three axes parallel to those of WCS. Based on the RCS, rotations of all intermediate joints can be performed for motion reconstruction purpose.

3.2. Implementation of MTP technique

To generate the best motion reconstruction from 2D monocular correspondence with the proposed MTP method, one key point is to find the most appropriate setting of the weighting parameters in each AF_i. Based on the attributes of residuals in AFs, the WP settings are divided into two categories: 2D WP setting and 3D WP value. WP setting for orientation and length of each segment, joint positions on 2D image plane belong to the first category, because such parameters are only related to the calculation of 2D residuals. It is obvious the WP of each joint's 3D recovered position in the previous frame should belong to the other category, which mainly concerns 3D attributes of human postures. As 2D residuals and 3D residual have different units, it is suggested to derive the possible 2D WP setting first and 3D WP value for rotations about each intermediate joint will be evaluated during run time based on the generated 2D WP setting.

As long as the recovered posture's projection is close to its corresponding image features, it appears similar to the original posture from the same viewpoint the video was produced. Actually there exist numerous possible postures which have such same projection, however not all of them resemble the original posture in all details. To establish the relationship between different 2D WPs in AFs, some simple motion sequences, which concern the human motion nearly parallel to the image plane, especially the translation of *pelvis*, are reconstructed. During these reconstructions all possibilities just mentioned are found out and the sum of residuals of each 2D factor is calculated. According to statistic data obtained, the possible 2D WP setting is finally determined as [21 20 10], following the order as segment's orientation, segment's length, and joint's position.

Once 2D WPs are obtained, the value of real-time 3D WP employed during rotations about the i^{th} intermediate joint in the $(K+2)^{th}$ frame can be derived through the following Eq. (9) and (10):

$$WP_K^{3D} = 15000 \quad (K = 2) \tag{9}$$

$$T_j^{K-1}(s) = \left[(1-s) * R_j^{K-1} + s * E_j^{K-1}\right]$$

$$WP_K^{3D} = average\left(\sum_{s=0}^{1} \frac{\left\{ \sum_{j=1}^{n} T_j^{K-1}(s) - \sum_{j=1}^{n} R_j^{K-2} \right\}}{\sum_{j=1}^{n} \left(P_{T_j^{K-1}(s)} - I_j^{K-1} \right)} \right)$$

$$(K \geq 3) \tag{10}$$

in the above formula s is a coefficient taking value from 0 to 1, while R_j^{K-1} and E_j^{K-1} are state vectors that respectively represent recovered and estimated 3D position of the i^{th} intermediate joint's direct descendant (the j^{th} joint) at the $(K\text{-}1)^{th}$ frame; thus a transition state $T_j^{K-1}(s)$ between R_j^{K-1} and E_j^{K-1} can be expressed as $\left[(1-s) * R_j^{K-1} + s * E_j^{K-1}\right]$ which depends on the value of coefficient s; $P_{T_j^{K-1}(s)}$ is the j^{th} joint's projection resulted from the transition state; finally I_j^{K-1} is the original 2D correspondence of the j^{th} joint at the $(K\text{-}1)^{th}$ frame.

4. Experiments and Statistics

In order to generate reliable motion reconstruction, 2D feature information (2D correspondence) extracted from the source images or video, such as each joint's position on the 2D image plane, must be highly accurate. There have been a large number of approaches to feature extraction in the image processing and computer vision area [1, 2, 7, 16]. However up to date, no technique is able to guarantee such process is already sufficiently accurate. Extraction error remains an unavoidable issue. As a matter of fact, such inaccuracy is one of the main factors preventing the progress of reconstruction technology. Besides, the lack of depth information in monocular image source makes it extremely difficult to evaluate the performance of any 3D reconstruction algorithm.

To enable accurate evaluation and assessment of the proposed MTP method, computer synthesized source videos are used as input before the MTP method is applied on real video data. The popular BVH (Bio-vision hierarchical data) motion files are firstly employed in any computer animation software to generate various animation series on a fixed human model with known geometry (for the time being, the motion is generated for a skeletal model to show the motion more clearly). Then the generated 3D animation is projected on the image plane to produce the accurate 2D posture correspondence in every frame, which is the only input to the motion reconstruction system. In this way errors in feature extraction can be eliminated. The 3D motion data and the camera settings are not utilized in any way during the reconstruction process. The 3D motion data is only used for comparison purpose after the 3D motion is reconstructed to ensure proper evaluation of the MTP method.

After the visual and numerical comparison based on synthesized input data, the MTP method is applied on real monocular videos to further test its accuracy. Since the depth information is not available in such cases, the technique can only be evaluated through the visual resemblance in the same viewpoint with the original, and the smoothness and reasonableness of the reconstructed motion in other viewpoints.

Currently only monocular sequences with resolution of 640x480 pixels and frame rate of 20fps are studied.

4.1. Results from "synthesized" data

In this section, the MTP technique is evaluated on computer synthesized monocular video data. The statistics data obtained during reconstruction from one sample kicking (22frames, where the joint *pelvis* is moving all the time) is shown in Fig.4, where the total 2D residuals of all the 17 joints between image and projection features are presented in form of stacked line. It can be seen from Fig.4 that, the total value of AFs of all joints at each frame for the kicking sequence reaches its peak at the 11th frame, which is only 1.660629 in pixels. The result indicates that the projection of the recovered 3D posture at each frame is very close to the original, given the image resolution of 640*480.

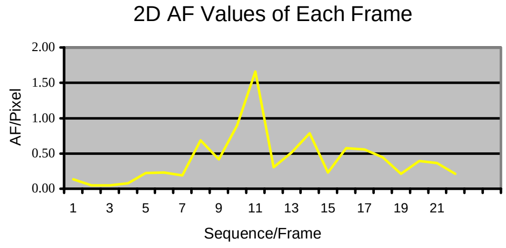

Figure 3: Total 2D AF value at all frames (Kicking)

Since the sequence is computer synthesized, the actual 3D position of every joint is available. In Fig.4 the original and reconstructed motions are still very similar when viewed from another angle.

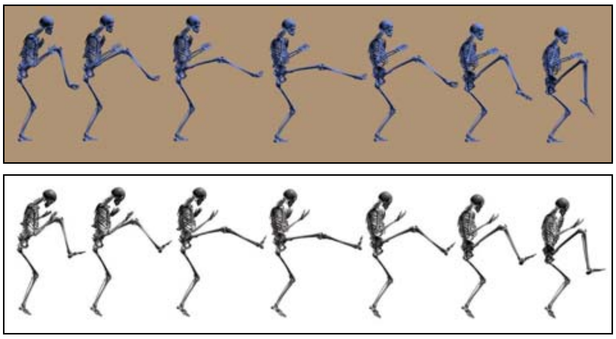

Figure 4: Input and reconstructed motion from other viewing directions

4.2. Results from real video data

Further the technique is tested on real monocular video sequences. One motion sequence composed of walking and squatting (49frames), and the reconstructed human animation based on it are presented in Fig.5. To ensure accurate image feature extraction from real video sequence and to reduce unnecessary noises, color labels are stuck to the human object at joint positions. Image processing techniques such as those mentioned in [2] are used to extract the 2D joint features from each frame of the video sequences; however we can only guarantee the noises will be minimized as possible as we can, which will affect the final 3D motion reconstruction.

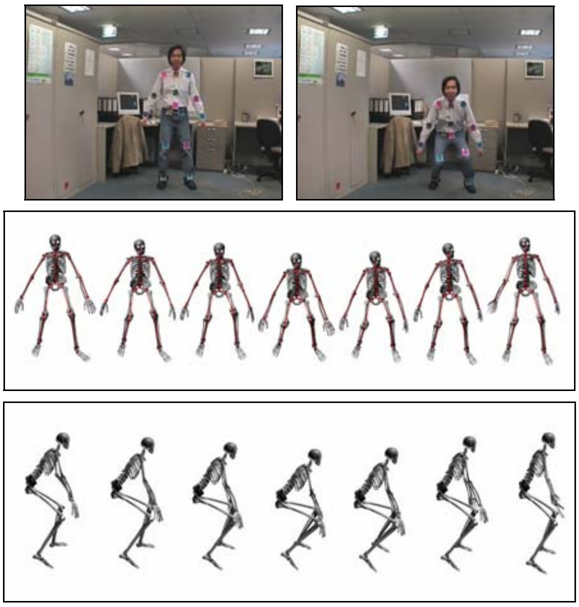

Figure 5: **Top:** Two frames from the input video **Middle:** Reconstructed motion - front view **Bottom:** Reconstructed motion - side view (Walking and Squatting)

Fig.6 shows the total values of AFs of all joints at each frame of this sequence. Comparing with Fig.3, it can be seen that the total values of AFs for real video reconstruction is much higher than those from "synthesized" data. However the maximum AF sum value is only 6.28112 at the 18th frame in a resolution of 640*480. The reconstruction from real video still can be considered as highly accurate.

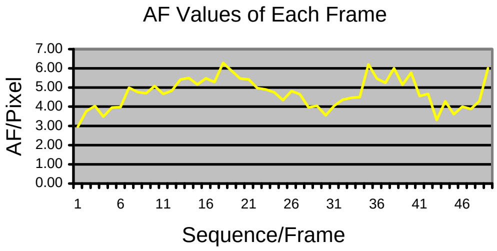

Figure 6: Total AF value at all frames (Walking and

squatting)

5. Conclusion

A model-based technique is proposed in this paper for motion reconstruction from un-calibrated monocular video sequences containing unrestricted human motion. MTP technique is first developed to reconstruct human motion with no body relocation. The technique is then extended to derive a new RCS at each frame, and hence enable reconstruction of any unrestricted human motion.

The main advantage of our approach is that through it a truly wide range of monocular sequences could be reconstructed efficiently, and there is no requirement for camera calibration. From experimental results presented in the paper, the reconstruction results are highly satisfactory as long as the 2D image features are reasonably accurate.

As a future work we are planning to introduce control on the leaf joint into MTP. Plausible motions about these parts are expected to be simulated.

Reference

[1] Agarwal A., Triggs B.: Learning to track 3D human motion from silhouettes. In *Proceedings of the 21st International Conference on Machine Learning* (July 2004), pp. 9-16.

[2] Barrón C., and Kakadiaris I. A.: A convex penalty method for optical human motion tracking. *First ACM SIGMM international workshop on Video surveillance*, IWVS'03, pp. 1-10.

[3] Callennec B. L., Boulic R.: Interactive motion deformation with prioritized constraints. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation* (Aug. 2004), pp. 163-171.

[4] David E. DiFranco, Tat-Jen Cham, James M. Rehg.: Reconstruction of 3D figure motion from 2D correspondences. In *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2001.

[5] Fdor M.: Application of inverse kinenatics for skeleton manipulation in real-time. In *Proceedings of the 19th spring conference on Computer graphics* (Apr. 2003), pp. 203-212.

[6] Favreau L., Reveret L., Depraz C., Cani, M. P.: Animal gaits from video. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation* (Aug. 2004), pp. 277-286.

[7] Gibson D. J., Oziem D. J., Daltion C. J., Campbell N. W.: Capture and synthesis of insect motion. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation* (Aug. 2005), pp. 39-48.

[8] Holt R. J., Netravali A. N., Huang T. S., Qian R. J.: Determining Articulated Motion from Perspective Views: A Decomposition Approach. *Pattern Recognition, Vol. 30* (1997), pp. 1435-1449.

[9] Hodgins J. K., O'Brien J. F., Bodenheimer R. E.: *Computer Animation*. In *Wiley Encyclopedia of Electrical and Electronics Engineering*, John G. Webster, ed., v. 3, 686-690, 1999.

[10] Ieronutti L., Chittart L.: A virtual human architecture that integrates kinematic, physical and behavioral aspects to control h-anim characters. In *Proceedings of the tenth international conference on 3D Web technology* (Mar. 2005), pp. 39-48.

[11] Kurihara T., Miyata N.: Modeling deformable human hands from medical images. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation* (Aug. 2004), pp. 355-363.

[12] Liu Xiaoming, Zhuang Yueting, Pan Yunhe.: Video based Human animation technique. In *Proceeding of the 7th ACM International Conference on Multimedia* (Oct. 1999), pp. 353-362.

[13] Park M. J., Choi M. G., Shin S. Y.: Human Motion Reconstruction from inter-frame feature correspondences of a single video stream using a notion library. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation* (July 2002), pp. 113-120.

[14] Remondina F., Roditakis A.: Human figure reconstruction and modeling from single Image or monocular video sequence. In *Proceeding of the 4th International Conference on 3D Digital Imaging and Modeling* (Oct. 2003), pp. 116-123.

[15] Taylor C. J.: Reconstruction of articulated objects from point correspondences in a single image. *Computer Vision and Image Understanding, Vol. 80, No. 3* (Dec. 2000), 349-363.

[16] Tsang W., Singh K., Fiume E.: Helping hand: an anatomically accurate inverse dynamics solution for unconstrained hand motion. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation* (Aug. 2004), pp. 319-328.

[17] WITKIN A., BARR A., FLEISCHER K.: Energy constraints on parameterized models. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques* (Aug. 1987), pp. 225-232.

[18] Zhao Jianhui, Li Ling.: Human motion reconstruction from monocular images using genetic algorithms. *Computer Animation and Virtual World, Vol. 15, No. 3-4* (July 2004), 407-414.