applied sciences

MDPI

Article

Motion Capture Research: 3D Human Pose Recovery Based on RGB Video Sequences

Xin Min [1], Shouqian Sun [1], Honglie Wang [1], Xurui Zhang [1], Chao Li [1] and Xianfu Zhang [1,2,*]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[2] School of Jewelry and Art Design, Wuzhou University, Wuzhou 543002, China
* Correspondence: zhangxianfu@zju.edu.cn

Received: 28 July 2019; Accepted: 23 August 2019; Published: 2 September 2019

Abstract: Using video sequences to restore 3D human poses is of great significance in the field of motion capture. This paper proposes a novel approach to estimate 3D human action via end-to-end learning of deep convolutional neural network to calculate the parameters of the parameterized skinned multi-person linear model. The method is divided into two main stages: (1) 3D human pose estimation based on a single frame image. We use 2D/3D skeleton point constraints, human height constraints, and generative adversarial network constraints to obtain a more accurate human-body model. The model is pre-trained using open-source human pose datasets; (2) Human-body pose generation based on video streams. Combined with the correlation of video sequences, a 3D human pose recovery method based on video streams is proposed, which uses the correlation between videos to generate a smoother 3D pose. In addition, we compared the proposed 3D human pose recovery method with the commercial motion capture platform to prove the effectiveness of the proposed method. To make a contrast, we first built a motion capture platform through two Kinect (V2) devices and iPi Soft series software to obtain depth-camera video sequences and monocular-camera video sequences respectively. Then we defined several different tasks, including the speed of the movements, the position of the subject, the orientation of the subject, and the complexity of the movements. Experimental results show that our low-cost method based on RGB video data can achieve similar results to commercial motion capture platform with RGB-D video data.

Keywords: 3D human pose recovery; motion capture; generative adversarial constraint; convolutional neural network

1. Introduction

Currently, in the digital video industry, motion capture methods based on high-speed cameras and key-point markers on the body surface are widely used [1]. However, this approach has the following disadvantages: (1) The markers on the human-body surface limits the freedom of human-body movement; (2) A controlled and restricted site is needed to obtain high-accuracy data; (3) Calculation is very time consuming, etc. On the other hand, the unmarked motion capture method based on ordinary optical cameras is convenient and inexpensive. In this method, the accuracy of motion capture is low and cannot meet some fields with high precision requirements, such as medical rehabilitation [2]. However, the precision requirements of the digital game industry and virtual human-machine ergonomics are not very strict, the rapid development of unmarked motion capture devices based on optical systems has been greatly promoted, and such devices have the benefits of acceptable computing time and convenient use.

The existing mainstream three-dimensional (3D) pose perception method consists of two parts: multi-camera [3–5] and monocular camera [6–8]. The solution of restoring 3D human poses from

monocular images is arbitrary and uncertain, and most methods can only obtain the 3D skeleton points of the human body, but cannot directly obtain the motion information such as the rotation matrix between the nodes [9,10], which needs to be solved by assumptions and inverse kinematics. To achieve the aforementioned goal, this paper proposes an end-to-end solution combined with a deep convolutional neural network, which can restore a 3D human grid model from single image data, and then smooth the serialized 3D human model using the correlation between RGB video frames and obtain the final human grid model. The solution proposed is to calculate the parameters of the parameterized skinned multi-person linear (SMPL) model [11] to restore the 3D pose of the human body. To constrain the arbitrariness in the solution process: On the one hand, we use 2D/3D skeleton point constraints to obtain a more accurate human-body model and combine human height constraints to reduce the arbitrariness problem in the two-dimensional (2D) to 3D solution process. On the other hand, we introduce the idea of a generative adversarial network (GAN) [12] and use the GAN constraint to generate a more human-like 3D human model. To evaluate the effectiveness of the motion capture method based on 3D human pose recovery, as proposed in this paper, we obtained the monocular-camera video sequences and depth-camera video sequences from the motion capture platform consisting of two Kinect (V2) devices and iPi Soft commercial software. Then we compared the performance of the proposed solution and this commercial motion capture platform in four different tasks for the variations of speed, position, orientation, and complex tasks. Experimental results show that the system performance of our low-cost method based on RGB (monocular-camera) video data is comparable to that of the commercial motion capture platform with RGB-D (depth-camera) video data.

1.1. Motion Capture

The current motion capture devices are mainly divided into the following five categories based on different technical principles: (1) Optical markerless motion capture devices, which usually use depth cameras or calibrated normal cameras to obtain 3D skeleton point coordinates of the human body based on RGB color images or depth images. This method has the advantages of easy deployment (e.g., iPi Soft), and being low priced. It has been widely used in motion capture of video games, and some scholars have used this platform as a data acquisition device for scientific research [1,13–15]. Modern functional evaluation and biomechanical research are mostly based on the use of optoelectronic systems. Ancillao et al. reviewed the working principle of this system, its application, and settings in the clinical practice [16]. Cappozzo et al. used optoelectronic stereophotogrammetric data to address the issues related to the 3D reconstruction of motion and analysis of skeletal system kinematics in vivo [17]; (2) Optical marker motion capture devices, which achieve motion capture with active or passive markers placed on key points of the body surface. This method is highly accurate but expensive (e.g., Vicon). Bevilacqua et al. utilized the "Vicon 8" motion capture system to capture the movements of a dancer in 3D and modified the system to control digital music [18]; (3) Mechanical motion capture devices, which perform attitude perception through mechanical wearable devices. This method is cheap, but the people move unnaturally because of the influence of mechanical equipment (e.g., Animazoo Gypsy). There is a wearable technology market for athletes and coaches, feedback from the device allows them to focus more on the movements they are performing and enables them to perform biomechanical changes without a coach [19]; (4) Magnetic motion capture devices, which calculate the position and orientation by the relative magnetic flux of the three orthogonal coils of the transmitter and receiver. This method is less expensive, but the return value at the boundary is not accurate due to nonlinear characteristics (e.g., Ascension MotionStar). Gawsalyan et al. reported a typical root mean square error (RMSE) of approximately 7° for the magnetic, angular rate and gravity (MARG) sensors for the detection of upper limb motion in cricket [20]; (5) Inertial motion capture devices, which collect the attitude orientation of body parts through the wireless action attitude sensor, and recover the human motion model based on the principle of human kinematics. This method is moderately accurate, suitable for outdoor activities with good portability, but the external wear sensor is easy to move, resulting in

measurement errors (e.g., Xsens). Low-power and low-cost electronic sensors have been successfully implemented through direct measurement by accelerometers, which have been used to continuously monitor patients in clinical and home environments [21]. Godfrey et al. investigated the use of a low-cost instrumenting gait with an accelerometer, and they observed that appropriate gait algorithms were an effective tool for estimating total steps and average spatiotemporal gait characteristics [22]. Mannini et al. discussed the computational algorithms for classifying human physical activity using on-body accelerometers [23]. Ancillao et al. identified, selected, and categorized the methodologies for estimating the ground reaction forces from kinematic data obtained by inertial measurement units worn by the subject, laying the foundation for kinetic analysis and motion performance testing outside the laboratory [24].

1.2. Human Pose Recovery

The human pose recovery based on visual perception can be mainly divided into three parts according to different outputs: (1) Human pose recovery based on 2D skeleton points, which can be divided into single-person pose estimation [25], and multi-person pose estimation [26] according to the number of human bodies; (2) Human pose recovery based on 3D skeleton points, which recovers the 3D skeleton points of the human body directly from pictures or videos. According to the different processes, they can be divided into the two-step method [27], and the direct method [28]; (3) Human pose recovery based on 3D human models, which estimates the parameters of the parametric human-body model, generally using the parameterized SMPL model [103]. The results of 3D skeleton points and 2D skeleton points of the human body can be obtained by the forward solution of the pose restoration of 3D human model. For this reason, it is more challenging than the previous two methods. Bogo et al. first proposed using 2D skeleton point constraints to solve SMPL parameters, and made many assumptions and manual constraints to make the generated human-body model more realistic [29]. Thiemo et al. recovered parametric human-body models from monocular video streams through human contours, including shape coefficients, contour coefficients, texture, and hair parameters [30]. Angjoo et al. proposed an end-to-end 3D human pose recovery method by using the idea of GAN instead of manual rule constraints [31] and our proposed approach was inspired by this.

1.3. Overview

To summarize, this paper proposes a pose-recovery approach of motion capture for recovering 3D human poses from unmarked RGB video sequences. This approach mainly includes three parts: end-to-end 3D human pose estimation based on single frame images, human-body pose generation based on video streams, and motion capture experiment verification based on 3D human pose recovery. The main contributions of this paper are as follows:

- The end-to-end method for 3D human pose estimation based on a single image is combined with 2D/3D skeleton point constraints and human height constraints to generate the 3D human model with higher precision. Meanwhile, the traditional method of manually defining rules is replaced by GAN constraints, which is used to generate a more human-like 3D human model;
- Combining the correlation of video sequences, a 3D human pose recovery method based on video streams is proposed, which uses the correlation between videos for smoothing to generate a more stable 3D human pose;
- Using two Kinect devices and iPi Soft series software to build a platform for motion capture experiments, an approach of using RGB-D video sequences is compared with the proposed approach of using RGB video sequences to verify the effectiveness of the proposed solution. In addition, experimental datasets are available to the public for related academic research.

2. Proposed Method

Our method is a composite of three main stages: 2D human pose estimation, 3D human pose estimation based on single frame images (Section 2.1), and human-body pose generation based on

video streams (Section 2.2). First, we adopt the open-source Openpose [32] framework as the method of 2D pose estimation, estimating the coordinates of 2D human skeleton points from the input color image, which is used as the projection constraints $L_{reproject}$ for 3D human pose estimation. Second, we take a 2D color image as input, combined with the projection constraints of 2D pose estimation and the human height constraint L_{height}, and use a code network architecture (encoder) combined with three-level neural network decoding (decoder) to obtain the deviation values of the SMPL model and camera parameters. Then, referring to the characteristics of the GAN, adversarial error $L_{adversarial}$ is used to constrain the authenticity of the generated human-body model. Last, we obtain a smoothed 3D human motion model, adding inter-frame continuity constraints L_{smooth} for continuity and correlation between video frames, and adding initial state constraints $L_{initial}$ for the accuracy of initial state, and finally extract the human skeleton in the sequence to obtain a 3D human pose. Figure 1 illustrates the whole pipeline of the proposed 3D human pose recovery approach, where the input is a video sequence containing human motion and human height, and the output is an 85-dimensional parameter, which can be calculated to obtain a 3D human pose skeleton.

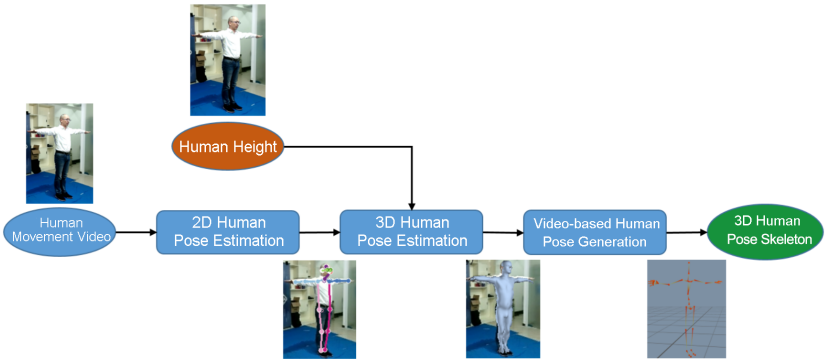

Figure 1. The pipeline of the proposed 3D human pose recovery approach.

2.1. 3D Human Pose Estimation Based on a Single Frame Image

The deep convolutional neural network is used to solve the SMPL model [11] parameters by the obtained 2D human pose and the input human height (H). This section includes the SMPL model, the end-to-end network structure, and the model pre-training process.

2.1.1. SMPL Model

This SMPL model contains 6890 vertices (N) and 23 key points (K). The initial state of this model contains the human-body average template: $\overline{T} \in \mathbb{R}^{3N}$ (T-pose initial pose) and the blend weight: $\mathcal{W} \in \mathbb{R}^{N \times K}$ (Figure 2a). The human-body shape blend deformation function of this model: $B_S(\vec{\beta}) : \mathbb{R}^{|\vec{\beta}|} \mapsto R^{3N}$, and human-body key-point prediction function: $J(\vec{\beta}) : \mathbb{R}^{|\vec{\beta}|} \mapsto \mathbb{R}^{3K}$, their inputs are human-body shape coefficients: $\vec{\beta} \in \mathbb{R}^{|\vec{\beta}|}$ (Figure 2b). Human pose blend deformation function of this model: $B_P(\vec{\theta}) : \mathbb{R}^{|\vec{\theta}|} \mapsto \mathbb{R}^{3K}$, the input is the axial angle of the human-body pose parameters: $\vec{\theta} \in \mathbb{R}^{3K}$ (Figure 2c). Finally, the standard linear blend skinned function: $W(\cdot)$ is used to drive the model deformation.

In summary, the SMPL model is defined as: $M(\vec{\beta}, \vec{\theta}, \Phi) : \mathbb{R}^{|\vec{\beta}| \times |\vec{\theta}|} \mapsto \mathbb{R}^{3N}$, where $\Phi = [\overline{T}, \mathcal{W}, \mathcal{S}, \mathcal{J}, \mathcal{P}]$ is the parameter obtained by solving the SMPL model, including: human-body average model, blend weight, orthogonal basis of human-body shape principal components: $\mathcal{S} = \left[\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_{|\hat{\beta}|}\right] \in \mathbb{R}^{3N \times |\vec{\beta}|}$, human-body key-point regression matrix: $\mathcal{J} \in \mathbb{R}^{3K \times 3N}$, and a vcpose blend deformation vector: $\mathcal{P} = [\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_{9K}] \in \mathbb{R}^{3N \times 9K}$. Through the method of grid alignment [33], the multi-pose dataset [11] and multi-shape dataset [34] obtained from 3D scanning are aligned with the grid of SMPL model, then the aligned dataset is used to train and solve the model parameters: Φ.

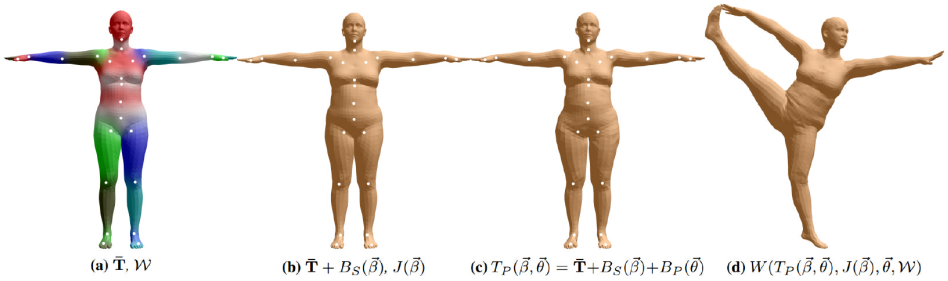

Figure 2. An illustration of the SMPL model.

2.1.2. End-to-End Network Structure

Using the weak-perspective camera model, we assume that the global rotation coefficient is $\boldsymbol{R} \in \mathbb{R}^3$ (indicator of the axial angle), the translation vector is $\vec{t} \in \mathbb{R}^2$, and the scaling coefficient is $s \in \mathbb{R}$. The SMPL model contains 23 joints, hence we can get rotation matrices by pose parameters $\vec{\theta}$, using the Rodriguez formula: $\exp\left(\overline{\omega}_j\right) = I + \hat{\overline{\omega}} \sin\left(\left\|\vec{\omega}_j\right\|\right) + \hat{\overline{\omega}}^2 \cos\left(\left\|\vec{\omega}_j\right\|\right)$, where $\vec{\omega}_j$ is the skew symmetric matrix of 3-vector, $\overline{\omega} = \frac{\vec{\omega}}{\|\vec{\omega}\|}$ and I are the 3×3 identity matrix. Therefore, combined with the shape coefficient solved by the SMPL model: $\vec{\beta} \in \mathbb{R}^{10}$, and the relative 3D rotation coefficient: $\vec{\theta} \in \mathbb{R}^{3*23}$ (indicator of the axial angle), we can obtain 85 parameters: $\Theta = \{\vec{\beta}, \vec{\theta}, \boldsymbol{R}, \vec{t}, s\}$ that need to be solved in the 3D pose recovery. On the one hand, the human pose parameters: $\vec{\theta}$, human shape parameters: $\vec{\beta}$, and extra global parameters: $\boldsymbol{R}$, $\vec{t}$, s are used to calculate the GAN constraint; On the other hand, we can get the 2D/3D human pose using the output parameters, hence the 2D/3D key-point constraints can be obtained. Similar to existing research [31], the global loss function is calculated as follows:

$$L = w_r L_{reproject} + 1_{3d} w_{3d} L_{3d} + w_a L_{adversarial} + 1_{height} w_h L_{height} \tag{1}$$

where 1_{3d}, 1_{height} are indicator functions, when the real label exists, the value is 1, otherwise the value is 0. The following are the different components of the loss function, where W represents the weight of the loss function, $L_{reproject}$ represents the error value between the key point of the 3D human model projected onto the 2D plane ($\hat{\boldsymbol{x}}$) and the real 2D key point ($\boldsymbol{x}$):

$$L_{reproject} = \sum_i \left\| v_i \left(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i \right) \right\|_1 \tag{2}$$

where $v_i \in \{0, 1\}$ indicates whether the key point of i is visible. The projection process of $\hat{\boldsymbol{x}}$ is calculated as follows:

$$\hat{\boldsymbol{x}} = s\Pi(\boldsymbol{R}\boldsymbol{X}(\vec{\beta}, \vec{\theta})) + \vec{t} \tag{3}$$

where Π represents an orthogonal projection, $\boldsymbol{X}(\vec{\beta}, \vec{\theta})$ represents the set of 3D key points calculated by SMPL model. $L_{reproject}$ uses the projection constraints of the 2D space, so that the generated 3D human model projection basically coincides with the key points of the 2D space.

The solutions of 3D models generated by the projection constraints are not unique, and some of the results may not be the real human body. So in order to avoid the diversity of solutions, the 3D constraint is calculated as follows:

$$L_{3d} = L_{3d\,\text{joints}} + 1_{smpl} L_{3d\,\text{smpl}} \tag{4}$$

$$L_{3d\,\text{joints}} = \sum_i \left\| \boldsymbol{X}_i - \hat{\boldsymbol{X}}_i \right\|_2^2 \tag{5}$$

$$L_{3d\,\text{smpl}} = \sum_i \left\| \left[\vec{\beta}_i, \vec{\theta}_i \right] - \left[\hat{\vec{\beta}}_i, \hat{\vec{\theta}}_i \right] \right\|_2^2. \tag{6}$$

The real 3D human parameters: $\vec{\beta}, \vec{\theta}$ can be calculated by the Mosh method [35] using real 3D skeleton points. 1_{smpl} is an indicator function, when the SMPL parameter exists, the value is 1, otherwise the value is 0.

Without any a priori condition constraints, the generated 3D human model may have some abnormalities, such as limb crossing, abnormal bending, and abnormal limb morphology. In the existing research [29,36], some a priori constraints are added, such as each part of the body is assumed to be a capsule and as far as possible away from the centers of these capsules to resolve the limb crossing, the elbow joint and the knee joint can only be bent in a fixed direction to solve the abnormal bending, the shape parameters conform to a fixed distribution to solve the abnormal limb morphology. There is a loss of precision in recovering 3D human model from 2D images with an error of 2–5 cm in height [30]. If the height information of human body is known in advance, the accuracy of the generated 3D human model can be improved, and the height loss function is calculated as follows:

$$L_{\text{height}} = \| f_h(\Theta) - height^* \|_2^2 \tag{7}$$

where f_h represents the function of calculating the human-body height through 3D human parameters, and *height** represents the real human-body parameter.

However, rule-based methods sometimes fail to fully consider various abnormal situations, and rule definitions are too complex to facilitate model training. To replace the traditional method of manually defining rules, we refer to [31,37,38] and combine the idea of GAN [12], using the classifier to distinguish between the real-mark human-body parameter model and the human-body parameter model generated by the generator, which makes the discriminator difficult to distinguish, so the human-body model conforming to the real structure can be generated through such an adversarial method. In our method, the GAN model is easy to converge, and the dimensions of the human parameters that need to be judged are low, so other types of GAN are not selected. As shown in Figure 3 , the discriminant network (D) is used as the data-driven rule constraint, so that the 3D human model parameters $\Theta = \{\vec{\beta}, \vec{\theta}, \boldsymbol{R}, \vec{t}, s\}$ generated by the generated network (G) are as close as possible to the real distribution, and the loss function is calculated as follows:

$$L_{adversarial} = L_G + L_D \tag{8}$$

$$L_G = \sum_i (D_i(G(I)) - 1)^2 \tag{9}$$

$$L_D = (D_i(\Theta^*) - 1)^2 + D_i(G(I))^2 \tag{10}$$

where i represents the number of different discriminators, and we use K + 2 discriminant networks for shape coefficients: $\vec{\beta}$, pose coefficients: $\vec{\theta}$, and each key point. I represents the input image. Θ^* represents the true 3D human parameters. The purpose of the generator loss function: L_G is to make the Θ generated by the generated network (G) close to the true value 1 in the output of discriminant network (D). The loss function of the discriminator: L_D can distinguish the real human-body parameters and the human-body parameters generated by the generator, to make the discrimination ability of the discriminant network (D) stronger.

Finally, the end-to-end network structure is explained (Figure 3). (1) For the encoding network, we utilize the pre-trained ResNet50 [39] on the ImageNet dataset [40], using the pooled 2048-dimensional features as the encoding network output; (2) For the regression network, we adopt a three-iteration network [41,42], and use iteration error feedback to obtain accurate output parameters, in which the input of each iteration is a splicing of 2048-dimensional encoding features and 85-dimensional current parameters, and the output is the deviation value $\triangle\Theta$ between 3D parameters and real parameters. The hidden layer of the regression network uses a two-layer 1024-dimensional full connection network, and the output layer is the number of cameras and 3D human model parameters ($|\Theta| = 85$); (3) For the discriminant network, we use K+2 fully connected discriminant networks:

The shape coefficient of the SMPL model, the number of input units is $|\overrightarrow{\beta}| = 10$; the discriminant network of the rotation matrix of human-body key points (K), the number of input units is nine; the discriminant network of the rotation matrix of human-body all key points, the number of input units is $32K$ (32 hidden layers of neurons in each key-point discriminant network are spliced). For the discriminant network of shape coefficients, the number of neural network units in the two hidden layers is divided into 10, 5; For the discriminant network of each key point, there is a full connection layer containing two layers of 32 neurons, and the parameters of K discriminant networks are shared. For the discriminant network of all key points, the input is a splice of 32 feature layers (the second full connection layer) for each key point, and the hidden layer is also a full connection network with 1024 neurons in two layers respectively. The number of output units of the $K + 2$ discriminant networks are all 1, and the range of values is [0, 1], indicating the probability that the input 3D parameters belong to the real human body. Through the GAN constraint in discriminant network to replace the traditional manual rule constraints, the 3D human parameters generated by the generated network can be driven by a large number of real 3D human-body data.

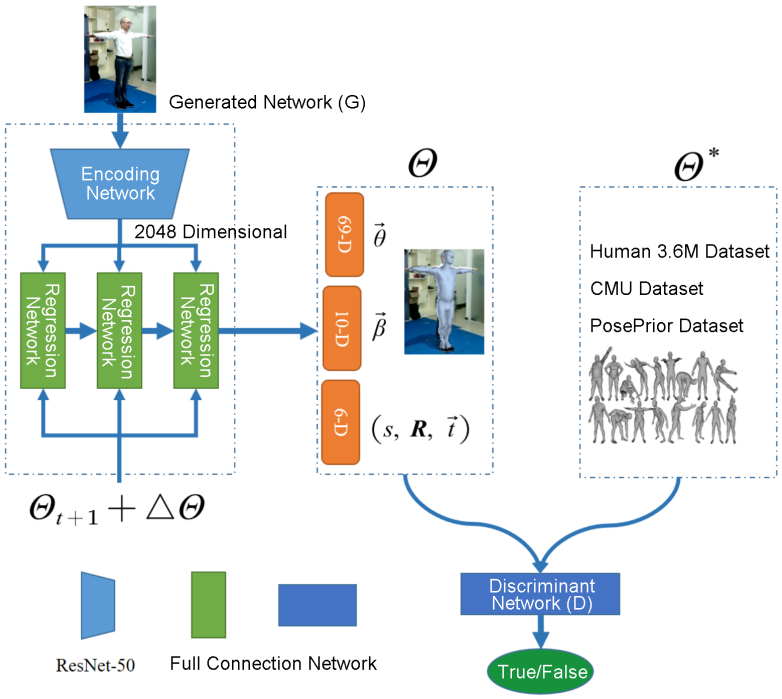

Figure 3. The end-to-end 3D human-body pose estimation based on single-frame images.

2.1.3. Model Pre-Training Process

In the training of end-to-end networks, two data sources are adopted in the training process to get the parameterized SMPL model. One is the data source of the generated network, providing image data and label data for the network, and the other is the data source of the discriminant network, providing only the parameters of the SMPL model. 2D key-point coordinates with real marks are required to calculate the 2D projection error ($L_{reproject}$), and the open-source Leeds Sports Poses (LSP) dataset, LSP-extended dataset [43], Motion Picture Industry Institute (MPII) dataset [44] and Microsoft COCO (MS COCO) dataset [45] are used. In calculating the error of 3D key points ($L_{3d\ joints}$), 3D key-point coordinates with real marks are required. We use the open-source 3.6 Million accurate 3D Human poses (Human3.6M) dataset [46], the Max Planck Institute-Informatics-3D Human Pose (MPI-INF-3DHP) dataset [47], and the iPi dataset obtained by the motion capture platform in this paper. In calculating the 3D parameter errors and SMPL real parameters of discriminant network, we use the Mosh method to obtain the SMPL parameters from real 3D key points, and calculate the loss of $L_{3d\ smpl}$ using only the Human3.6M dataset [46]. In training the depth discriminator network, we use datasets with SMPL

parameters including: the Human3.6M dataset, the Carnegie Mellon University (CMU) motion capture dataset, and the PosePrior dataset [48]. The datasets used for network training are summarized in Table 1. When calculating the 2D/3D key-point constraints, the inconsistency of the number of key points in the dataset will affect the calculation of 3D human key points. Therefore, the key points of all datasets are converted according to the order of the key points in the LSP dataset, and the key points that do not exist are treated as 0. For the MS COCO dataset, five key points of the face are added, and the numbers and names of the converted key points are shown in Figure 4.

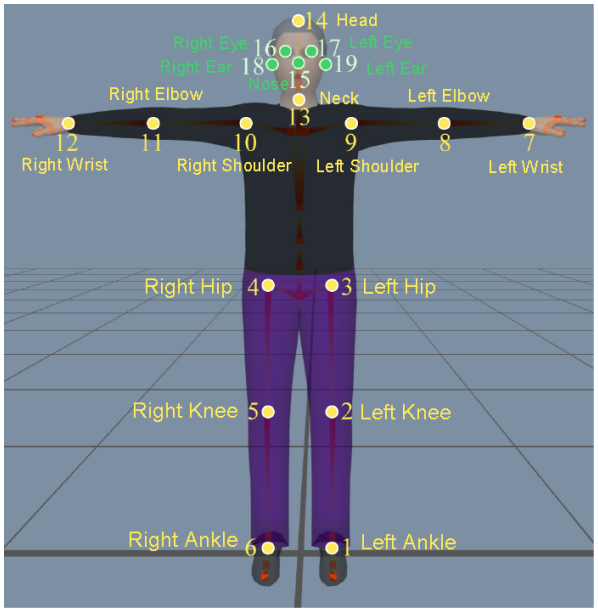

Figure 4. The numbers and names of the converted key points (14/19 skeleton points).

Table 1. Overview of datasets for network training. LSP: Leeds Sports Poses; MPII: Motion Picture Industry Institute; CMU: Carnegie Mellon University.

Constraint Types	Dataset	Training Sample	Key Point	Converted Key Point	Loss Function
2D Constraint	LSP	1000	14	14	$L_{reproject}$
	LSP-extended	10,000	14	14	$L_{reproject}$
	MPII	20,000	16	14	$L_{reproject}$
	MS COCO	79,344	17	14 + 5	$L_{reproject}$
3D Constraint	Human3.6M	312,188	24	14	$L_{3d\ joints}$, $L_{3d\ smpl}$
	MPI-INF-3DHP	147,221	28	14	$L_{3d\ joints}$
	iPi Dataset	**20,955**	**65**	**14**	$L_{3d\ joints}$, L_{height}
Discriminant Constraint	Human3.6M	1,559,985	-	-	$L_{adversarial}$
	CMU	3,934,267	-	-	$L_{adversarial}$
	PosePrior	181,968	-	-	-

In the pre-training process, we use the Rodriguez formula to convert the pose parameters from the generated network $\vec{\theta}$ into a 3×3 rotation matrix, which is used as the loss function and the input of the discriminant network. The learning rate of the generated network is set to 0.0001, the learning rate of the discriminant network is set to 0.00001, and the weight decay coefficient of the generated network and the discriminant network are both 0.0001. Using the skeleton point annotation of the image in each dataset, the matrix diagonal of the human body is scaled to 150 pixels, and the image is scaled to 224×224 size as the network input, while maintaining the aspect ratio of the image. In order to increase the diversity of training samples, random scaling, translation, and flip horizontals are used in the preprocessing of input images, in which the threshold for translation is set to [0, 20],

and the threshold for random scaling is set to [0.8, 1.25]. The super parameters are selected through the experiments of validation set, the weights of loss functions with different constraints are as follows: $w_r = 60$, $w_{3d} = 60$, $w_{height} = 100$, $w_a = 1$.

The open-source framework we use for pre-training is Tensorflow [49] and the computing hardware is NVIDIA's Titan XP. To balance the proportion of true/false samples in the discriminant network, the batch size of the data source of generated network is set to 64, and the batch size of the data source of discriminant network is set to 3×64. In the training process, Adam algorithm [50] is firstly used to iterate 50 rounds to accelerate convergence, and the stochastic gradient descent (SGD) algorithm [51] is used to iterate 10 rounds to get more accurate results.

2.2. Human-Body Pose Generation Based on Video Streams

The end-to-end pose generation method is only for the input of a single image. When the input is a video sequence, this method cannot consider the correlation between the connected frames, resulting in a large jitter in the generated 3D model sequence. In addition, the video sequence can solve the problem that human-body parts are occluded in a certain frame. So taking advantage of video sequences during the inference phase will be able to generate a smoother and jitter-free parameterized SMPL model.

The process of human-body pose generation based on a video stream is shown in Figure 5. The encoding network is known, and the three-iteration regression networks in the generated network (G) are referred to as the decoding network. In the smooth solution, we use the method of optimizing the 2048-dimensional feature space obtained by the encoding network $\boldsymbol{Z} = \{\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_N\}$. Specifically, the network parameters after the pre-training are kept constant, and the feature space $\boldsymbol{Z}$ obtained by the encoding network is directly optimized. In this way, the feature of the video sequences can be obtained by using forward propagation, and then the feature space of the entire video sequence is used as the parameter to be optimized, and the loss function is inversely propagated for optimization. So that this solution process only needs to be imported into the decoder network, which is small in scale and can optimize the feature space of about 1000 pictures at a time.

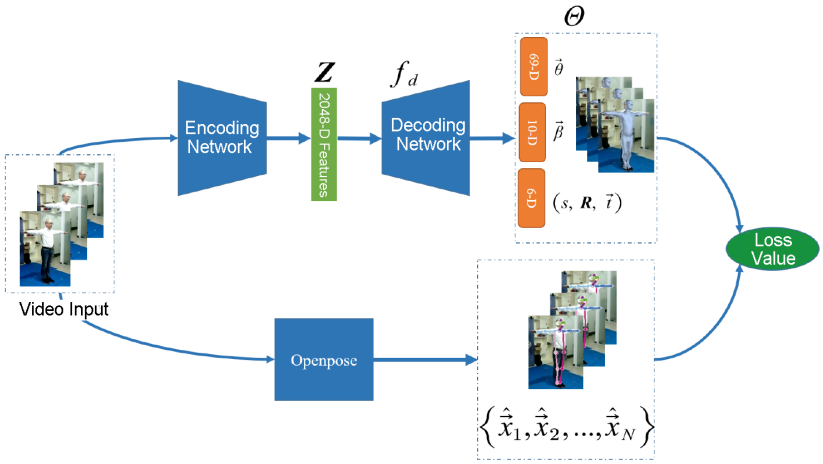

Figure 5. The process of human-body pose generation based on video sequences.

First, Openpose is used to extract 2D key points in each frame $\left\{\hat{\vec{x}}_1, \hat{\vec{x}}_2, \ldots, \hat{\vec{x}}_N\right\}$, N is the number of video frames. The most intuitive 2D skeleton point constraint is calculated as follows:

$$loss_{2d} = \sum_t \sum_j c_{t,j} ||\hat{\vec{x}}_{t,j} - \Pi \left[J_j \left(f_d \left(\vec{z}_t\right)\right)\right]||_1 \tag{11}$$

where t is the serial number of each frame, j is the serial number of each key point, f_d is the decoding function, J is the human-body key-point prediction function, which can get the 3D skeleton point coordinates through the 3D human-body parameters, Π is the orthogonal projection, which uses

camera parameters to project 3D key points onto a 2D plane. The 2D key points extracted by Openpose have been very accurate to some extent. By adding this 2D skeleton constraint, the feature space can be further optimized to produce a more accurate 3D human model.

Then, the end-to-end pre-training model is trained with a large amount of marked data, and the obtained 3D human-body sequence has jitter, but basically conforms to the normal human-body distribution. Therefore, when further optimizing the solution, some useful feature information of the 3D skeleton points in the initial state is retained, and only the inaccurate part is optimized. The initial 3D skeleton point constraint is calculated as follows:

$$loss_{3d\ \mathrm{init}} = \sum_t w_t \left(\left\| \hat{\vec{\theta}} - [f_d(\vec{z}_t)]_\theta \right\|_2^2 \right) \tag{12}$$

where $\hat{\vec{\theta}}$ is the 3D pose model of the initial state obtained by the pre-training model, and is also the converted rotation matrix 9×9. w_t is the weight used to maintain the useful information in $\vec{z}_t$, and optimize the useless information, the calculation process is as follows:

$$w_t = \exp(-\delta_t) \tag{13}$$

$$\delta_t = \sum_j c_{t,j} \left\| \left(\hat{x}_{t,j} - \Pi J_j(f_d(\vec{z}_t)) \right) \right\|_2^2. \tag{14}$$

Here, the validity of the information in $\vec{z}_t$ is measured by the 2D key-point error. The smaller the 2D key-point error, the larger the weight w_t, so the decoded feature space should be closer to the initial state.

After, by calculating the error of 3D key points between adjacent frames as a constraint condition, the result is smoothed and the feature space is optimized. The smoothness constraint is calculated as follows:

$$loss_{\mathrm{smooth}} = \sum_t \sum_j \left\| J_j(f_d(\vec{z}_t)) - J_j(f_d(\vec{z}_{t+1})) \right\|_2^2 . \tag{15}$$

Last, to obtain a more accurate feature space, we obtained human height information in the T-pose from our motion capture platform and the human height constraint is added. The human height constraint is calculated as follows:

$$loss_{\mathrm{height}} = \| f_h(f_d(\vec{z}_{\mathrm{tpose}})) - height^* \|_2^2 . \tag{16}$$

To sum up, the global constraint function is calculated as follows:

$$loss = w_{2d}\, \mathrm{loss}_{2d} + w_{3d\mathrm{init}}\, \mathrm{loss}_{3d\mathrm{init}} + w_{sm} loss_{smooth} + 1_h w_h loss_{height}. \tag{17}$$

where 1_h is an indicator function, when the image and human height of T-pose exist, the value is 1, otherwise the value is 0. Different weights represent different importances. In the calculation process, the weighs are set to: $w_{2d} = 10$, $w_{3d\ \mathrm{init}} = 100$, $w_{sm} = 25$, $w_h = 25$. Then the SGD is used to calculate the feature space $\boldsymbol{Z}$. The batch size is set to 1000, iteration is set to 100 times, and the learning rate is set to 0.001.

2.3. Results and Analysis

This section first verifies the effect of different constraints on the 3D human pose perception accuracy, then compares the video-based solution with the image-based solution.

2.3.1. Effect of Different Constraints on the 3D Human Pose Perception Accuracy

In the pose recovery based on a video stream, we added 2D skeleton point constraints, initial 3D skeleton point constraints, smoothness constraints, and human height constraints. We tested the

importance of different constraints to prove the rationality of the proposed constraint conditions. To exclude interference from other disturbing factors (e.g., speeds, occlusions), the slow and simple basic action of elevating right arm laterally with slow speed was repeated. Taking the RGB video data from the Kinect on the left as input, the experimental results are shown in Figure 6. The pink curve is the experimental result of our proposed solution, the purple curve is the reference result of the iPi Soft system, and the green curve below is the absolute value of the difference between the result obtained by our solution and the reference configuration. We also used traditional means to get some ground-truth measure, the two lines represent the right hand vertical position with maximum elevation (yellow) and lowest elevation (blue). Thus, we can find some macroscopic errors in body tracking of iPi Soft system because the purple curve exceeds the yellow line in some cases, this implies that iPi Soft system can only be used as a reference result and is not absolutely accurate. The three data in the table are the maximum, mean, variance of the absolute value of the deviation (in cm).

The results show that all the proposed constraints are helpful to improve the accuracy of motion capture. When one of the constraints is removed by the one-row method, the average error rises respectively: 24.3% (2D skeleton point constraint), 3.6% (initial 3D skeleton point constraint), 8.9% (smoothness constraint), 10.5% (human height constraint). The 2D skeleton point constraint has the greatest influence on the precision loss, this indicates that the 2D key points extracted by Openpose have high precision. The smoothness constraint increases the accuracy by 8.9%. This is because during the data acquisition process, some of the collected image data is ambiguous, but the previous and next frames may have higher quality. And the jitter of the error curve is severe without the smoothness constraint.

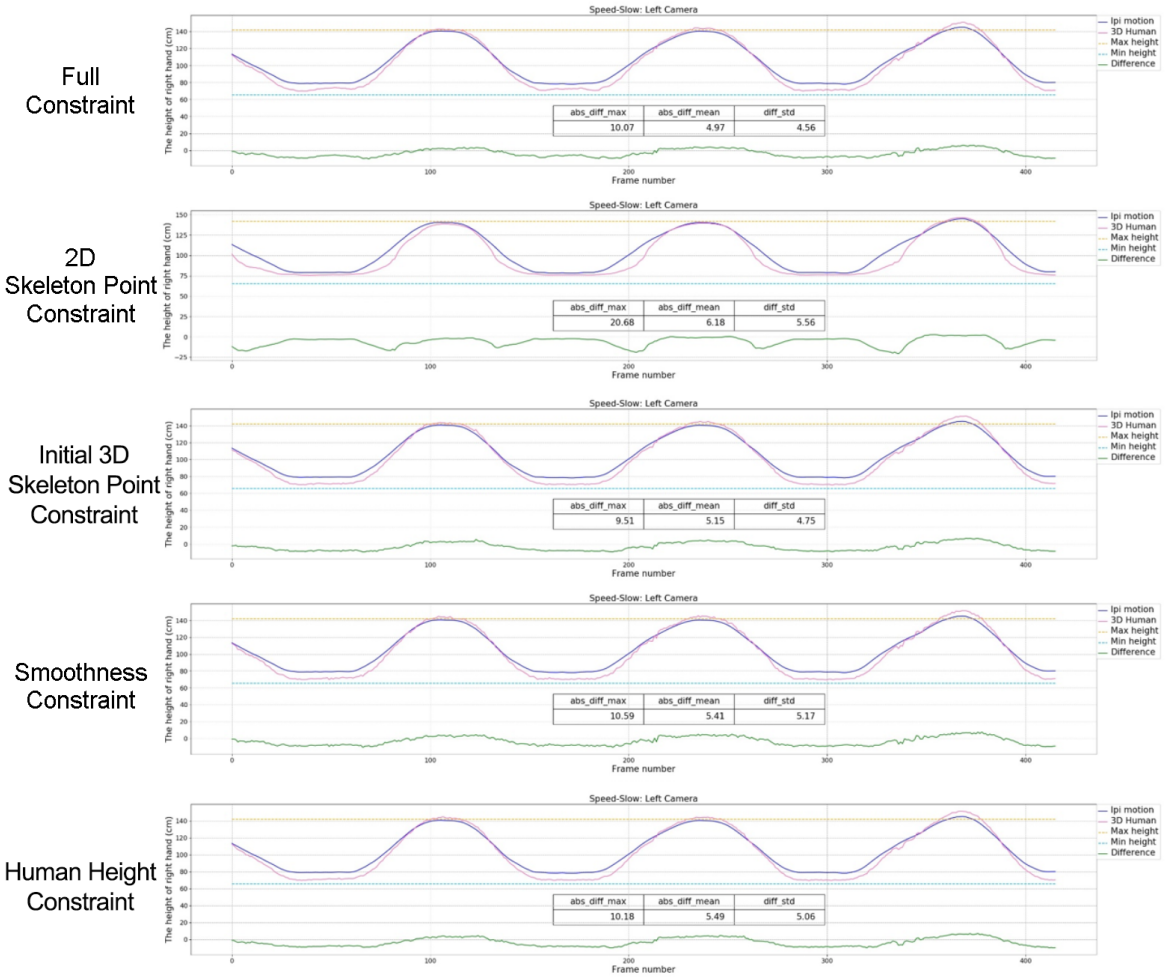

Figure 6. The effect of different constraints on the 3D human pose perception accuracy.

2.3.2. Validation of Human Pose Perception Based on Video Sequences

In this experiment, we compared the pre-trained model based on single image in Section 2.1 with the improved solution based on video stream in Section 2.2. The same as the previous experiment, the subject also repeated the side elevation of the right arm, the experimental results are shown in Figure 7. It can be seen from the results that the average error of the video-based solution on the left and right cameras is reduced by 22.7% and 41.7%, respectively, which also greatly reduces the instability and jitter of the results.

Since it is difficult to measure the absolute center position of two Kinect devices in the actual situations, there are some differences between the left and right camera results. The average error of the left camera is lower, and the accuracy of the experimental results is higher; the variance of the right camera is lower, and the experimental results are more stable. Therefore, we weighted the human pose parameters of the two cameras (weights were 0.5), and the results shown in Figure 8 were obtained. The experimental results show that the weights of left and right cameras can complement each other to improve the accuracy of motion capture. In particular, the results of the highest and lowest points of the right wrist are better than the reference results obtained by iPi Soft. Thus, in the following experiments, weighted results of left and right cameras are taken as the final results of the solution.

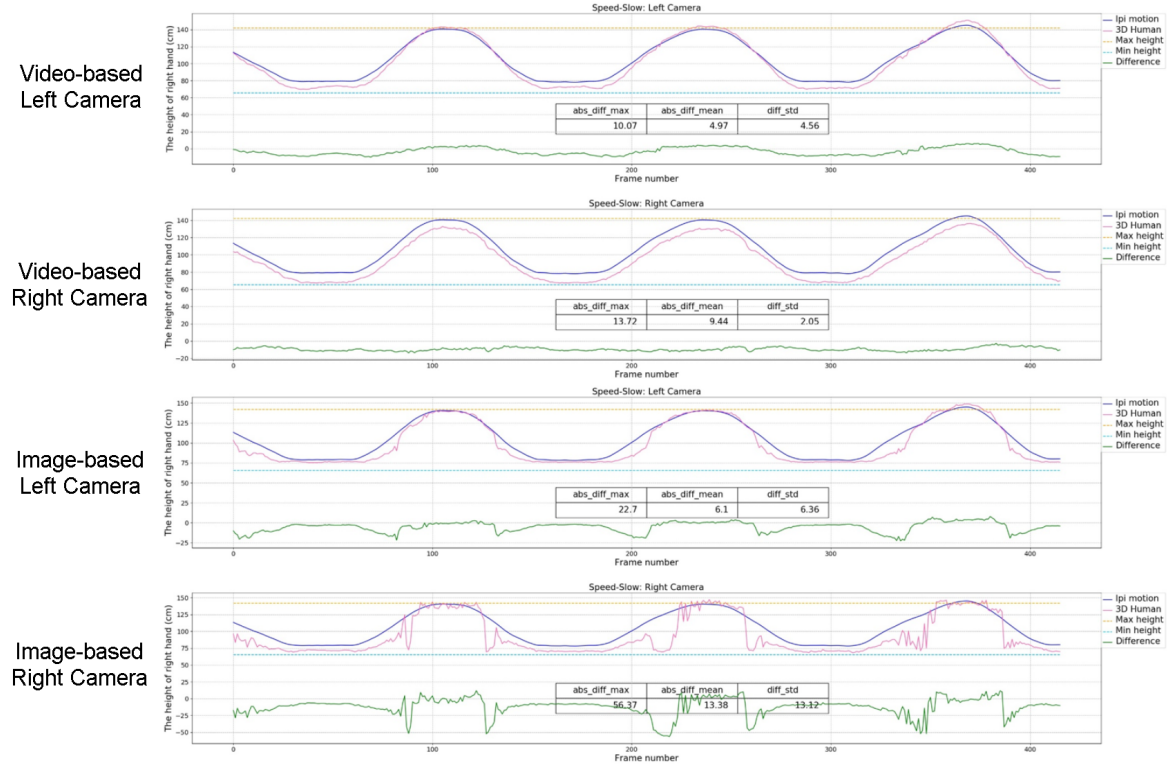

Figure 7. The comparison of the video-based solution and the image-based solution.

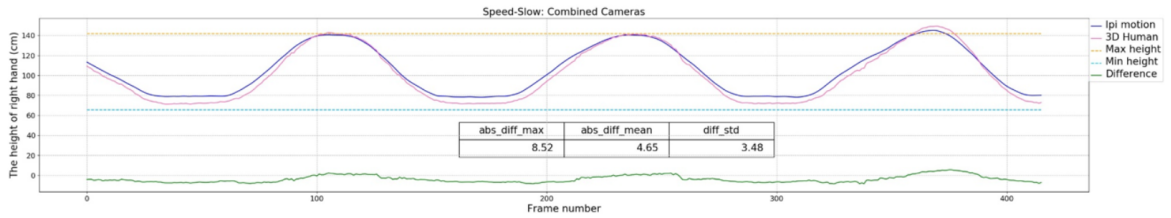

Figure 8. The weighted results of left and right cameras.

3. Experiment Implementation and Result Discussion

To verify the effectiveness of the proposed 3D human pose recovery solution in motion capture applications, we built a motion capture platform using two Kinect devices and iPi Soft series software, and used it as a comparison platform for experiments. This experiment was mainly used to analyze the accuracy of the algorithm in four different tasks: different movement speeds, different standing positions, different orientations, and complex motions. For these tests, we defined a set of basic movements of arms and legs, and full-body complex movements. We finally compared the performance of the proposed video-based solution and the commercial motion capture platform in these four scenarios, and verified the validity and convenience of the proposed solution.

3.1. Experimental Setup and Environment

The system is mainly composed of a calculator, two Microsoft Kinect 2.0s and two software modules. All of our experiments were conducted on two Kinects connected to a calculator. Kinect sensors are capable of recording standard RGB video and RGB-D video simultaneously. Therefore, the comparison is not performed at the sensor level. The two software modules are the recorder and the processor. The recorder we adopted is iPi Recorder, which is used to synchronize video from different sensors. The processor is iPi Mocap Studio, which is used to track body segments in video recording.

The two Kinect configuration need to be calibrated to configure the sensors' positions in the virtual scene. We recorded a video with magnesium light spiraling down in the acquisition area and detected the position of the light spot in each picture to calculate the external parameters of two devices. Once the calibration software showed "Good", we saved the calibration file for subsequent experiments, and the following step was to record movements. We recalibrated the device whenever we restarted the experiment. Video recorded was analyzed with iPi Mocap Studio and each key-point movement captured was exported with Biovision Hierarchy (BVH) files. The hierarchical structure of the human model can be analyzed with the BVH file containing rotations of each joint for each time step.

The disposition of sensors around the scene has been considered as well. Because different sensors have different sensory fields, the motion capture platform construction of iPi Soft system is generally determined by the choice of sensors and the kind of movements to be recorded. The maximum measuring distance of Kinect (V2) is 4 meters (m), and when the distance from the device exceeds 2.1 m, the maximum height range obtained is 2 m, which can basically cover the height range of most Asians. Meanwhile, we referred to the instructions of the iPi Soft system, a cross combination on the acquisition area could be suitable to acquire all movements of a standalone subject, which is a relevant tradeoff solution between performance and applicability. As shown in Figure 9, the two Kinects are placed on a 60–90° angle of horizontal direction with a distance of 2.5–3 m to the intersection of the two Kinects, and the area covered by each other is the movement area of the subject. Figure 10 shows the real experimental scene, we used two tripods to fix the height of the Kinect to 1 m. Since Kinect's depth information is generated by infrared pulses, we need to avoid: (1) direct natural light; (2) reflective materials in the experimental area. So we drew all the curtains and put two non-reflective carpets on the floor to improve tracking capabilities (Figure 10b). Moreover, the room light must be uniform and preferably from above, and the subject was not allowed to wear shiny fabrics.

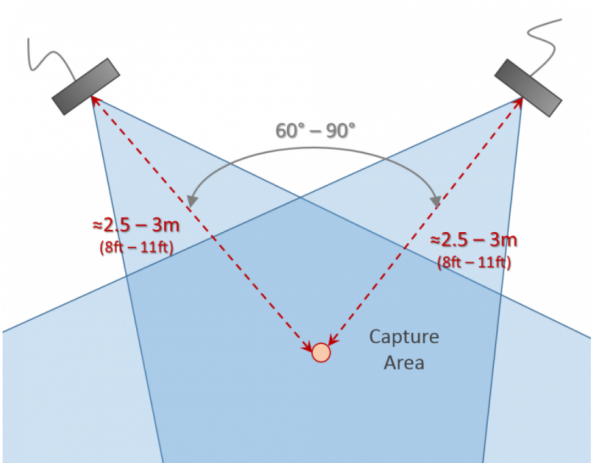

Figure 9. The motion capture platform construction with two Kinects.

(**a**) (**b**)

Figure 10. The real experimental scene. (**a**) Overall scene; (**b**) Acquisition area.

3.2. Types of Tasks

To verify the effectiveness of the proposed 3D human pose recovery solution to the field of ergonomics, biomedical and health on workplaces, and provide references for specific application scenarios with different accuracy requirements, we designed a series of tasks to explore different motion capture accuracies of this solution in execution speed, position in acquisition area, orientation relative to the sensors, and complexity of movements, which are the key elements of the movements that could influence data quality (Figure 11).

We asked the subject to perform a right-arm movement: (1) elevating the right arm to the shoulder height on the side; (2) at a constant speed; (3) repeating three times. To ensure the repeatability, we let the subject watch a demo video before the experiment. According to the iPi Soft requirements, the subject needed to maintain a T-pose for five seconds before any action begins. Execution speed can influence motion capture accuracy due to the limited frame rate of Kinect sensors. The subject: (1) faced straight ahead; (2) in the center of the acquisition area; (3) with different speed limits for slow and fast, respectively. Position and orientation parameters can change the image quality of Kinect sensors. For experiments on spatial position, the subject: (1) faced straight ahead; (2) at a slow speed; (3) changing position in the center to 50 cm on the left, on the right, on the front, and on the rear. For experiments on movement orientation, the subject: (1) in the center; (2) at a slow speed; (3) changing orientation in the center along with four directions with an angular displacement of 90°.

In some complex movement scenes, body parts are usually occluded and the overlapping body segments can affect the way Kinect sensors see the subject and the capability to track the subject. Therefore, to evaluate the performance of our systems more fully, four complex body movements were carried out: crossing arms, tying shoes, jumping, and walking, which can be illustrated by the skeletal view of the key position (e.g., head, hand, foot) in Figure 11. In addition, we recruited 60 subjects

to participate in the experiment of walking and manually segmented the gait data of each person according to each gait cycle, as the iPi dataset for model pre-training data.

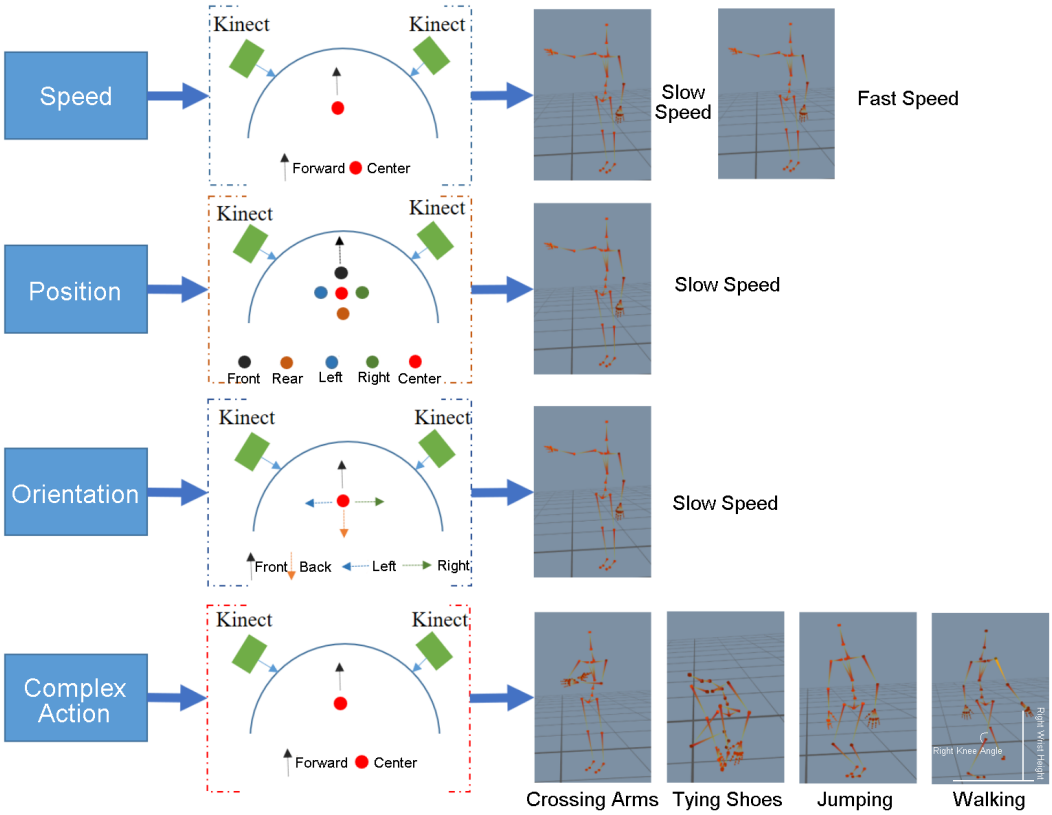

Figure 11. Different test tasks of the defining movements.

3.3. Results and Discussion

This section shows the results gained with RGB video sequences to RGB-D video sequences for tasks as described in the previous section, and also discusses the most interesting results with some comments. To verify the effectiveness of the proposed method, we manually measured the distance between the right wrist and the horizontal ground [13], and the angle of the right knee joint to highlight potential issues. We not only considered the key-joint positions of hands or feet as quantitative measures [52], but also visualized reconstruction results for qualitative analysis.

3.3.1. Speed of the Movements

Limited by the frame rate of the input device (e.g., Kinect), different speeds can affect the imaging effect. When the movement speed is too fast, a greater displacement is obtained between each frame, which can cause potential loss of information, and affect some vision-based algorithms. We tracked the task of elevating right arm at slow and fast speed respectively. Figure 12 shows the results of our proposed solution, the reference results of iPi Soft system, and their differences. Each graph also shows the maximum, mean, and variance of the deviation between the two acquisitions. The results indicate that the proposed solution is less affected by the speed and basically consistent with the reference curve, which can solve the image quality problem caused by speed. However, this solution overestimates the position of the hand, causing some fluctuations at the peak of the curve, especially during fast movements. This deviation is within an acceptable limit.

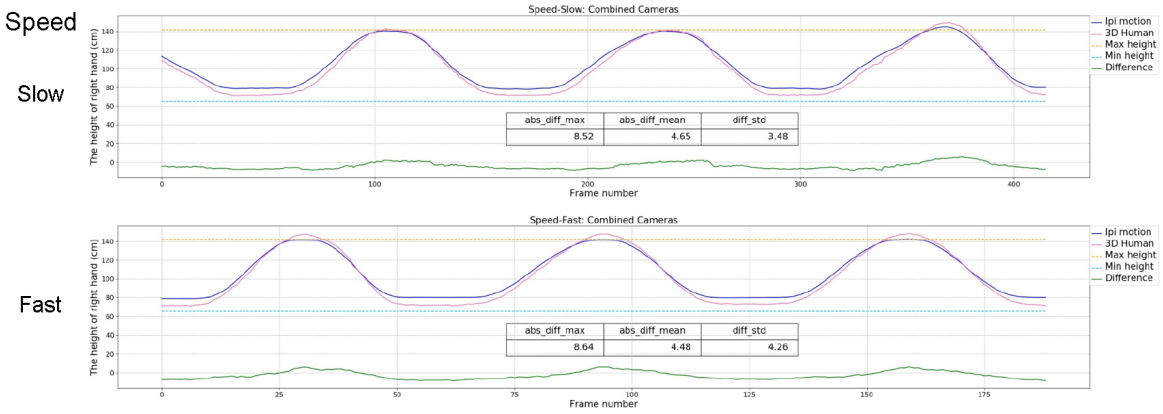

Figure 12. Results for the speed variations.

3.3.2. Position of the Subject

Measuring the position variations of movements has the aim of assessing the effect of position on motion capture accuracy. We selected four positions in the four directions of the front, rear, left, and right, 50 cm from the center point, and changed the subject position while repeating the movements of elevating the right arm. We can observe from Figure 13 that there are no significant differences between different positions, but close to the acquisition device within a certain range helps to improve the accuracy of motion capture. The smallest mean deviation is 2.55 cm (front) and the largest mean deviation is 5.17 cm (right). The results show that the system performance of the proposed solution is comparable to that of the iPi Soft system and even better than the reference results at some peaks.

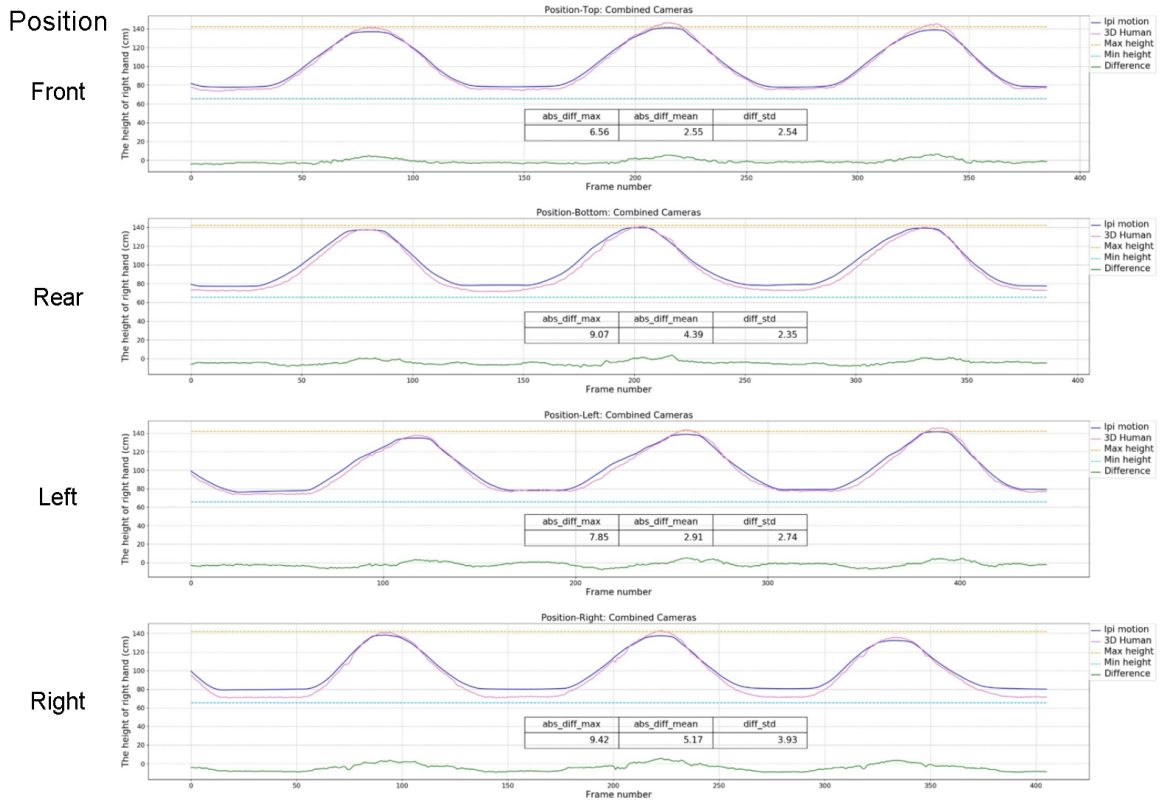

Figure 13. Results for the position variations.

3.3.3. Orientation of the Subject

When the subject moves in different orientations, the collected human image data varies greatly. For this test we defined four orientations: with the center point facing the two devices as the reference direction (front), the left, right, and back three orientations were selected at the angular displacement of 90°. The subject was also asked to repeat the right-arm movement. Figure 14 suggests that there are no substantial differences in different orientations, the maximum mean deviation from the reference curve is 4.97 cm (back) and the minimum is 2.53 cm (right). We also found that the performances in the same orientation (vertical/horizontal) are almost the same, and the experimental results in the vertical direction (left/right) are better than the horizontal orientation (front/back), where the mean deviation is twice higher. Because the image carries obvious depth information in the vertical orientation, which helps to improve recognition accuracy.

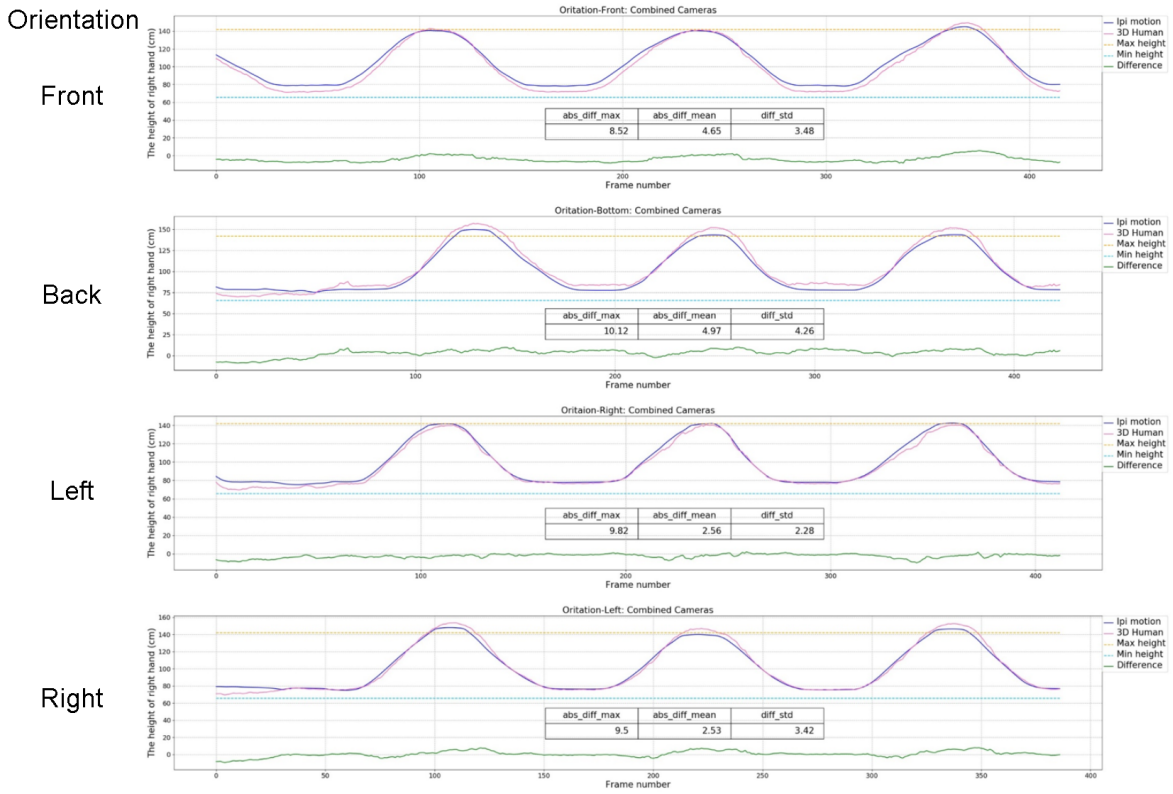

Figure 14. Results for the orientation variations.

3.3.4. Complexity of the Movements

The movement of the right arm was used to evaluate measures in all previous experiments. To verify the effectiveness of our proposed solution in more complex movements, we defined four complex action types: crossing the arms on the chest, kneeling down as to tie a shoe, two feet jumping, and walking back and forth. These tasks involve interference with overlapping body segments, changing perspective, intense movement, and loss of contact with the floor by one or both feet. All these movements reproduce common situations in everyday life, new retail, and human-computer interaction.

The experimental results are shown in Figure 15. All actions are still measured by the height of the right wrist from the ground, with the exception of walking measured by the angle of the right knee joint. It is found that the mean distance deviation between the right wrist and the ground distance of the proposed method does not exceed 5 cm, and the mean angular deviation of the knee joint does not exceed 5°. There are still some problems in the case of walking, the maximum angular deviation is 21.52°. This is because during the walking process, the right leg is completely overlapped. For the

iPi Soft system, it is also not tracked correctly. This indicates that when the subject is fully visible in complex movements there is no problem with tracking it. In addition, the differences between the results of complex tasks and elevating arm tasks are not significant.

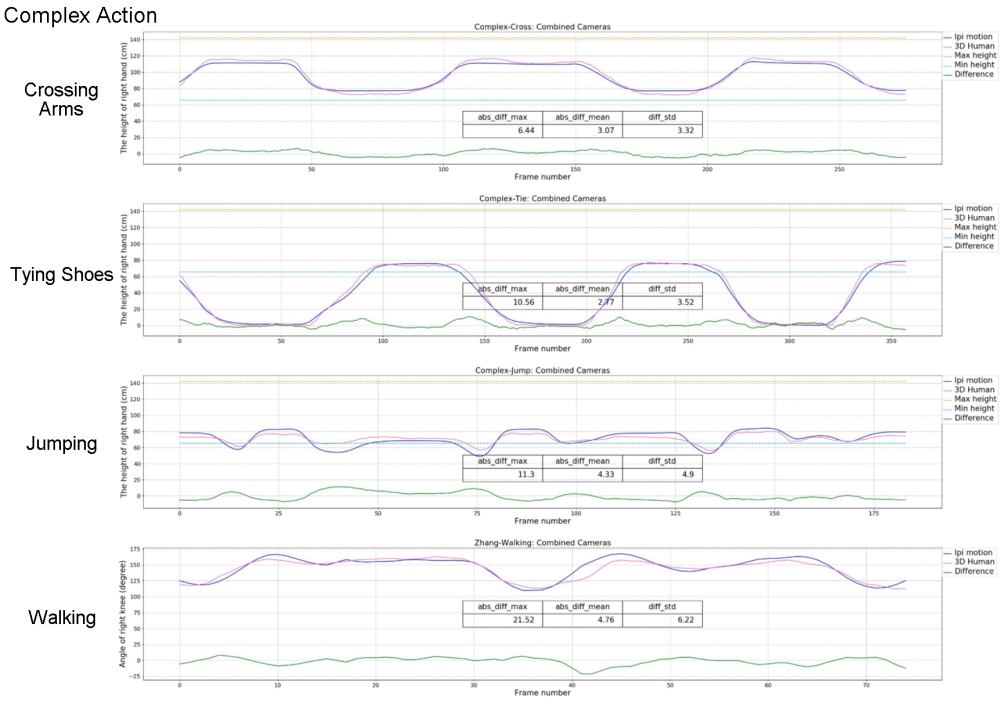

Figure 15. Results for the complex tasks.

To more fully demonstrate the effectiveness of our proposed solution, we also visualized the results of 3D pose recovery (Figure 16), which displays the plot of the most interesting body movements with the SMPL models. It can be seen from the visualization results that the solution based on video streams, proposed by us, can well recover the 3D human poses in the case of overlapping, coming close to the body, and losing contact with the ground.

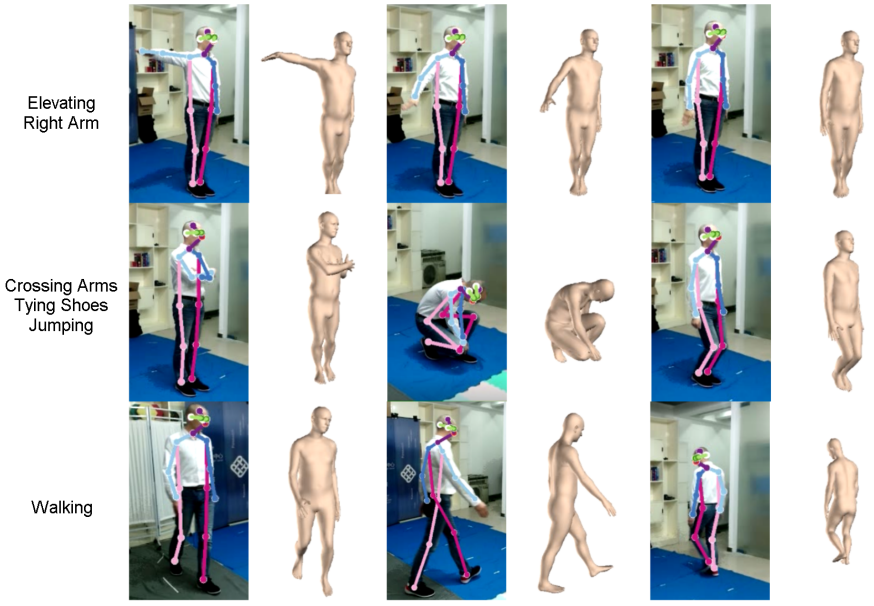

Figure 16. Visualized results of 3D human pose recovery.

4. Conclusions

In this paper, a 3D human pose recovery solution based on video sequences is proposed, and the SMPL model is used to model the 3D human body. Meanwhile, an end-to-end convolutional neural network is used to calculate the human-body pose parameters, shape parameters, and global camera parameters. In the process of 3D human pose perception, the first step is to pre-train the attitude-aware network based on a single image, using the 2D key-point dataset, 3D pose dataset, 3D human parameters dataset, and iPi dataset with human height information obtained by our motion capture experiment. The large-scale pre-training data ensures the network has a strong prediction and fitting ability. Meanwhile, the idea of GAN is introduced, which replaces the traditional prior rule constraints on the 3D human model, and simplifies the rule definition process, so that the generated 3D human model is as close as possible to the true structure distribution. In the second step, the video-based 3D human pose recovery method is further proposed. It is found that the feature space of deep convolutional neural networks can be directly optimized by integrating the correlation between video frames, the consistency of the initial state, and the human height constraint. Then, the feature space is decoded, so that the 3D human pose with higher precision is obtained by indirect calculation, and about 1000 picture sequences can be processed at a time. In the process of verifying the effectiveness of our proposed solution, we used two Kinect devices combined with the iPi Soft series software to build a motion capture platform and designed different motion capture experiments to compare and analyze the proposed solution with the self-built motion capture platform, at different speeds, different positions, different orientations, and complex actions. It is found that the inexpensive and convenient solution based on RGB video sequences proposed in this paper, combined with human height information, can obtain experimental results matching with commercial motion capture platform bases on RGB-D video sequences. The proposed method is beneficial to some key applications: The digital media industry (e.g., video games, animation, film), biomedical and health workplaces, and virtual human-machine ergonomics, which have very important, practical value.

Author Contributions: S.S. and X.Z. supervised the work; X.M. and X.Z. conceived and designed the experiments; C.L. and H.W. analyzed the data; X.M. wrote the paper.

Funding: This work was supported in part by the National Natural Science Foundation of China (grant #91748127) and National Key Research and Development Program of China (grant #2017YFD0700102).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3D	three dimensional
SMPL	skinned multi-person linear
GAN	generative adversarial network
SGD	stochastic gradient descent
RMSE	root mean square error
MARG	magnetic, angular rate and gravity
LSP	Leeds Sports Poses
MPII	Motion Picture Industry Institute
MS COCO	Microsoft COCO
Human3.6M	3.6 Million accurate 3D Human poses
MPI-INF-3DHP	Max Planck Institute-Informatics-3D Human Pose
CMU	Carnegie Mellon University
BVH	Biovision Hierarchy

References

1. Skals, S.L.; Rasmussen, K.P.; Bendtsen, K.M.; Andersen, M.S. Validation of musculoskeletal models driven by dual Microsoft Kinect Sensor data. In Proceedings of the International Symposium on 3d Analysis of Human Movement, Lausanne, Switzerland, 14–17 July 2014.
2. Colombo, G.; Facoetti, G.; Regazzoni, D.; Rizzi, C. A full virtual approach to design and test lower limb prosthesis. *Virtual Phys. Prototyp.* **2013**, *8*, 97–111. [CrossRef]
3. Hofmann, M.; Gavrila, D.M. Multi-view 3D Human Pose Estimation in Complex Environment. *Int. J. Comput. Vis.* **2012**, *96*, 103–124. [CrossRef]
4. Elhayek, A.; Aguiar, E.D.; Jain, A.; Tompson, J.; Theobalt, C. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
5. Hasler, N.; Rosenhahn, B.; Thormählen, T.; Michael, W.; Gall, J.; Seidel, H.; Informatik, M. Markerless motion capture with unsynchronized moving cameras. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
6. Chen, X.; Yuille, A. Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. *Eprint Arxiv* **2014**, *27*, 1736–1744.
7. Wang, C.; Wang, Y.; Lin, Z.; Yuille, A.L.; Gao, W. Robust Estimation of 3D Human Poses from a Single Image. In Proceedings of the Computer Vision & Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
8. Andriluka, M.; Roth, S.; Schiele, B. Monocular 3D pose estimation and tracking by detection. In Proceedings of the Computer Vision & Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
9. Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.P.; Xu, W.; Casas, D.; Theobalt, C. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *Acm Trans. Graph.* **2017**, *36*, 44. [CrossRef]
10. Peng, X.B.; Kanazawa, A.; Malik, J.; Abbeel, P.; Levine, S. *SFV: Reinforcement Learning of Physical Skills from Videos*; SIGGRAPH Asia 2018 Technical Papers; ACM: New York, NY, USA, 2018.
11. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: a skinned multi-person linear model. *Acm Trans. Graph.* **2015**, *34*, 248. [CrossRef]
12. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Bing, X.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
13. Regazzoni, D.; De Vecchi, G.; Rizzi, C. RGB cams vs RGB-D sensors: Low cost motion capture technologies performances and limitations. *J. Manuf. Syst.* **2014**, *33*, 719–728. [CrossRef]
14. Finn, M.T.; Smith, C.L.; Nash, M.R. Open-ended measurement of whole-body movement: A feasibility study. *Quant. Methods Psychol.* **2018**, *14*, 38–54. [CrossRef]
15. Han, S.U.; Achar, M.; Lee, S.H.; Penamora, F.A. Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring. *Vis. Eng.* **2013**, *1*, 6. [CrossRef]
16. Ancillao, A. *Modern Functional Evaluation Methods for Muscle Strength and Gait Analysis*; Springer: Berlin, Germany, 2018.
17. Cappozzo, A.; Croce, U.D.; Leardini, A.; Chiari, L. Human movement analysis using stereophotogrammetry: Part 1: theoretical background. *Gait Posture* **2005**, *21*, 186–196. [PubMed]
18. Bevilacqua, F.; Naugle, L.; Dobrian, C. Music control from 3D motion capture of dance. In Proceedings of the CHI 2001 for the NIME Workshop, Washington, DC, USA, 1–2 April 2001.
19. Adesida, Y.; Papi, E.; Mcgregor, A.H. Exploring the Role of Wearable Technology in Sport Kinematics and Kinetics: A Systematic Review. *Sensors* **2019**, *19*, 1597. [CrossRef] [PubMed]
20. Gawsalyan, S.; Janarthanan, T.; Thiruthanikan, N.; Shahintha, R.; Silva, P. Upper limb analysis using wearable sensors for cricket. In Proceedings of the 2017 IEEE Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 22–24 February 2017; pp. 1–6.
21. Godfrey, A.; Conway, R.; Meagher, D.; Olaighin, G. Direct measurement of human movement by accelerometry. *Med. Eng. Phys.* **2008**, *30*, 1364–1386. [CrossRef] [PubMed]
22. Godfrey, A.; Din, S.D.; Barry, G.; Mathers, J.C.; Rochester, L. Instrumenting gait with an accelerometer: A system and algorithm examination. *Med. Eng. Phys.* **2015**, *37*, 400–407. [CrossRef] [PubMed]

23. Mannini, A.; Sabatini, A.M. Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers. *Sensors* **2010**, *10*, 1154–1175. [CrossRef] [PubMed]
24. Ancillao, A.; Tedesco, S.; Barton, J.; Oflynn, B. Indirect measurement of ground reaction forces and moments by means of wearable inertial sensors: A systematic review. *Sensors* **2018**, *18*, 2564. [CrossRef] [PubMed]
25. Wei, S.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4724–4732.
26. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv* **2018**, arXiv:1812.08008.
27. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2659–2668.
28. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Honolulu, HI, USA, 21–26 July 2017; pp. 1263–1272.
29. Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.V.; Romero, J.; Black, M.J. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 561–578.
30. Alldieck, T.; Magnor, M.A.; Xu, W.; Theobalt, C.; Ponsmoll, G. Video Based Reconstruction of 3D People Models. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8387–8397.
31. Kanazawa, A.; Black, M.J.; Jacobs, D.W.; Malik, J. End-to-End Recovery of Human Shape and Pose. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7122–7131.
32. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
33. Bogo, F.; Romero, J.; Loper, M.; Black, M.J. FAUST: Dataset and evaluation for 3D mesh registration. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
34. Robinette, K.M.; Blackwell, S.; Daanen, H.; Boehmer, M.; Fleming, S. *Civilian American and European Surface Anthropometry Resource (CAESAR)*; Final Report; Technical Report; Sytronics Inc.: Dayton, OH, USA, 2002; Volume 1.
35. Loper, M.; Mahmood, N.; Black, M.J. MoSh: motion and shape capture from sparse markers. *ACM Trans. Graph. (TOG)* **2014**, *33*, 220. [CrossRef]
36. Zhou, X.; Sun, X.; Zhang, W.; Liang, S.; Wei, Y. Deep Kinematic Pose Regression. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 186–201.
37. Tung, H.F.; Harley, A.W.; Seto, W.; Fragkiadaki, K. Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation from Unpaired Supervision. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4364–4372.
38. Zheng, N.; Wen, J.; Liu, R.; Long, L.; Dai, J.; Gong, Z. Unsupervised Representation Learning with Long-Term Dynamics for Skeleton Based Action Recognition. In Proceedings of the 2018 Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 2644–2651.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
40. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.S.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
41. Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; Li, S.Z. Face Alignment Across Large Poses: A 3D Solution. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 146–155.
42. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human Pose Estimation with Iterative Error Feedback. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

43. Johnson, S.; Everingham, M. *Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation*; University of Leeds: Leeds, UK, 2010; pp. 1–11.
44. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In Proceedings of the Computer Vision & Pattern Recognition, Columbus, OH, USA, 24–27 Junauary 2014.
45. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014.
46. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
47. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In Proceedings of the International Conference on 3d Vision, Qingdao, China, 10–12 October 2017; Volume 271, pp. 506–516.
48. Akhter, I.; Black, M.J. Pose-conditioned joint angle limits for 3D human pose reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1446–1455.
49. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: a system for large-scale machine learning. In Proceedings of the Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
50. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
51. Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*; Physica-Verlag HD: Heidelberg, Germany, 2010.
52. Arai, K.; Asmara, R.A. 3D Skeleton model derived from Kinect Depth Sensor Camera and its application to walking style quality evaluations. *Int. J. Adv. Res. Artif. Intell.* **2013**, *2*, 24–28. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).