Hindawi
Security and Communication Networks
Volume 2023, Article ID 9894212, 1 page
https://doi.org/10.1155/2023/9894212

Retraction

Retracted: Deep-Learning-Based Motion Capture Technology in Film and Television Animation Production

Security and Communication Networks

Received 11 July 2023; Accepted 11 July 2023; Published 12 July 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Wei, "Deep-Learning-Based Motion Capture Technology in Film and Television Animation Production," *Security and Communication Networks*, vol. 2022, Article ID 6040371, 9 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 6040371, 9 pages
https://doi.org/10.1155/2022/6040371

Research Article

Deep-Learning-Based Motion Capture Technology in Film and Television Animation Production

Yating Wei

Wuhan University of Communication, Wuhan 430205, Hubei, China

Correspondence should be addressed to Yating Wei; weiyating@whmc.edu.cn

Received 3 December 2021; Accepted 19 January 2022; Published 11 February 2022

Academic Editor: Chin-Ling Chen

With the popularity of King Kong, Pirates of the Caribbean 2, Avatar, and other films, the virtual characters in these films have become popular and well loved by audiences. The creation of these virtual characters is different from traditional 3D animation but is based on real character movements and expressions. An overview of several mainstream motion capture systems in the field of motion capture is presented, and the application of motion capture technology in film and animation is explained in detail. The current motion capture technology is mainly based on complex human markers and sensors, which are costly, while deep-learning-based human pose estimation is becoming a new option. However, most existing methods are based on a single person or picture estimation, and there are many challenges for video multiperson estimation. The experimental results show that a simple design of the human motion capture system is achieved.

1. Introduction

Motion capture technology is more mature and common in the film and television industry. After capturing the motion data of professional actors, doing specific processing, and then binding with the character model in the film and television works, we can get 3D virtual animation [1–3]. The currently used pose capture system is mainly divided into two categories: sensor capture and optical capture. The former is more mature, characterized by fast transmission speed and more accurate pose data; the disadvantage is the higher cost, and wearable devices are less convenient to use. In contrast, optical capture is the opposite, and there are two types of optical capture: unmarked and marked. The object of this paper is a markerless capture system, where a common 2D image or video is used as input to capture the human body's joint point data using target detection and feature extraction [4]. Although it is not yet widely used due to its unstable performance, its advantages such as ease of use, flexibility, and low cost should not be overlooked.

In recent years, numerous scholars at home and abroad have proposed considerable convolutional neural network models and other auxiliary methods for human pose estimation, covering single to multiperson, 2D to 3D, and picture to video [5]. However, human pose is a complex nonlinear model, and environmental noise, occlusion, and spatial depth ambiguity are the main hindrances to this task. If the input object is video data, it is also a difficult task to output a high frame rate and smooth and stable pose. Most of the existing methods are based on image-based 3D pose estimation [6], or for single-person videos [7]. The actual motion capture application objects are many times facing multiperson scenes, and the characters must have contact with the virtual physical space, so we propose a 3D multiperson estimation model from the video to meet the practical needs.

There are two general types of 3D pose estimation: one is regressive—regressing the 3D coordinates of the nodes directly from 2D data, which require the data to be 3D labeled, which is often difficult to obtain—and the other is the lifting type [8] where the two-dimensional pose is first obtained and then a mapping method is trained to lift it to the three-dimensional space on top of the two-dimensional one. The work in this paper will be centered on the lifting style, where current 2D estimation methods are relatively mature. And we will focus on the implementation of 3D estimation especially for multiperson targets. Time-domain convolution makes full use of key information at different time points of the video stream to infer 3D pose. Treating the human key point connection relationship as a graph structure is a prerequisite for implementing graph convolution, and then, the 3D relationship of bones can be extracted from global and local together. Previous approaches focus on feature learning and large-scale data training on pixel space, without making good use of a priori information such as human kinematics, spatial-physical relationships, and human topology [9], and for monocular 3D pose, inference relying on neural network learning alone is not sufficient. Some common problems, such as spatial relative position errors of multiple targets, contact with the ground with penetration and vacillation, unnatural tilt of pose and mutual occlusion, are found from the practical use of some models.

The remaining sections of this article are arranged as follows: Section 2 describes the modern mainstream motion capture system; Section 3 is the main content of this article is the design of film and television animation based on motion capture technology; Section 4 is the multiperson 3D estimation network; Section 5 is the experiment and evaluation; Section 6 is the conclusion.

2. Modern Mainstream Motion Capture System

The current mainstream motion capture systems can be divided into four categories: mechanical, electromagnetic, acoustic, and optical [10]. Optical-based motion capture systems mainly use multiple cameras to capture motion image sequences and trajectories and then accomplish the task of motion capture by identifying and tracking specific markers in the image information and using the motion information of these marker points to perform the 3D reconstruction.

2.1. Mechanical Motion Capture System. Mechanical motion capture system relies on mechanical devices to track and measure the motion trajectory. Mechanical motion capture systems generally consist of multiple joints and rigid connecting rods [1]. When the device is in motion, the position and trajectory of the rod end point in space can be derived from the angle change measured by the angle sensor and the length of the linkage.

2.2. Acoustic Motion Capture System. Acoustic motion capture system consists of a transmitter, a receiver, and a processing unit. The transmitter is a fixed ultrasonic generator, and the receiver consists of 3 ultrasonic probes arranged in a triangular pattern [11, 12]. This type of device is relatively low cost, but the capture of motion has a large delay and lag, real time is poor, the accuracy is generally not very high, and the sound source and the receiver cannot have large obscuring objects between, by noise and multiple reflections and other interference. Since the speed of sound waves in the air is related to air pressure, the corresponding compensation must also be made in the latter algorithm.

2.3. Optical Motion Capture System. Motion capture system is the most widely used and convenient system in the world. It uses multiple infrared cameras to capture objects from different angles. The software is then used to analyze the image coordinates of the marker points on the image, and the 3D reconstruction is performed using computer vision principles to derive the motion data of the marker points. The advantages of optical motion capture are a large range of performer activities, no cable, mechanical device limitations, easy to use, and high sampling rate [13].

3. Design of Motion Capture-Based Animation for Film and Television

Motion capture technology needs to build a skeleton model in order to describe the motion of the real human body, and all the skeleton models we store in the motion module database are shown in the figure below, and in order to let the captured motion data to drive the 3D human model, we need to combine the model with the captured motion data to achieve matching with the model, so as to drive the movement of the model. Finally, the model is matched with the captured data, and the model can follow the captured motion data to move, as shown in Figure 1.

In the past, the movie was to make the real into virtual, but nowadays, it is making the virtual scene into reality [14]. A film and television works using motion capture technology only need the following steps can quickly produce a film and animation works liked by the audience, and the actual operation steps are shown in Figure 2.

We use motion capture technology for film, television, and animation production, which can greatly improve the level of film, television, and animation production. It can greatly improve the efficiency of film and animation production, reduce the cost of film and animation production, and make the film and animation production process more intuitive and more vivid effects [11, 15–17].

4. Multiperson 3D Estimation Network

The current single-person 3D estimation and multiperson 2D estimation methods are relatively mature, while multiperson 3D estimation has many challenges. The study in [18] proposes a multiview approach, which has good results but requires specialized datasets and is difficult to collect data from everyday scenes. In this paper, we use monocular video data to achieve multiperson motion capture at a lower cost. Current time-domain convolution and graph convolution are two important methods to achieve pose estimation, and

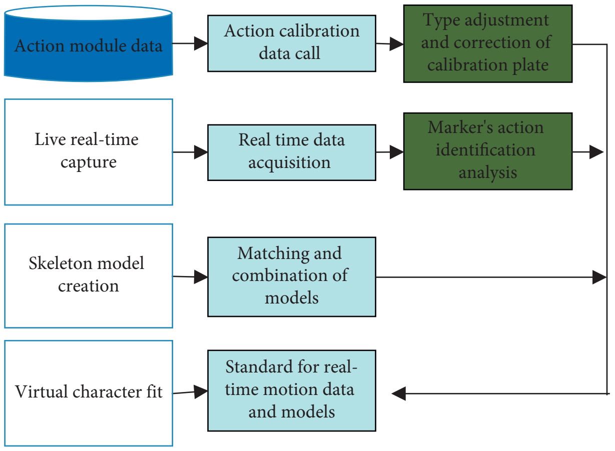
Action module data
Action calibration data call
Type adjustment and correction of calibration plate
Live real-time capture
Real time data acquisition
Marker's action identification analysis
Skeleton model creation
Matching and combination of models
Virtual character fit
Standard for real-time motion data and models

Figure 1: Motion capture-based film and television animation production process.

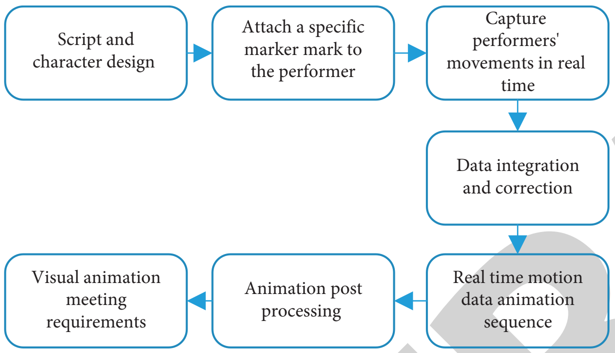
Script and character design
Attach a specific marker mark to the performer
Capture performers' movements in real time
Data integration and correction
Real time motion data animation sequence
Animation post processing
Visual animation meeting requirements

Figure 2: Motion capture technology for film and television production process.

combining the advantages of these two methods has also become a popular research direction [19]. Different targets in the real world have different distances from the camera, which is reflected in the image that the targets occupy different pixels, so the 3D distribution of multiple targets needs to be inferred from the 2D space, and the method used in this paper is depth estimation, and feeding the depth information to the time-domain graph convolution can achieve multiperson 3D estimation, which is the difference between multiperson and single-person.

4.1. Time-Domain Map Convolution.

The time-domain convolution network proposed in [20] exploits the continuity of the post in the time domain and performs a void convolution in the time domain from the two-dimensional pose input to lift it to the three-dimensional space. Time-domain convolution is derived from RNN and LSTM. The study in [21] avoids the fact that RNN networks cannot be processed in parallel in the time domain, and the gradient length between the input and output is fixed so that the input sequences of different lengths can be trained stably without gradient vanishing and exploding. The network also employs null convolution to obtain long time-domain information without being limited to the top and bottom frames. The model works well for 3D estimation in the video but is not applicable to multiperson situations. Inspired from [22, 23], the application of graph convolution to time-domain convolution enables to obtain better spatial information of joint points. The form of skeletal joint point data is a graph structure, and the graph convolution methods before [24] basically learn the weight parameters of the convolution layer, while for the complex nonlinear transformation from two-dimensional key points to three-dimensional space, understanding the spatial relationship between joint points and the influence of each joint point on other joint points is another focus that cannot be ignored. Human skeletal dynamics is a nonregular and complex structure, and 3D pose estimation using graph convolution can overcome this limitation, and the model extracts both local and global pose information to accommodate occlusion situations. The local and global optimization is shown in Figure 3. Since the human pose in the video stream has continuity with both local dynamic and local static as well as temporary occlusion needs to rely on both the auxiliary of up and down frames and the graph convolution for pose space estimation, the graph convolution and time-domain convolution are performed simultaneously in the network structure proposed in this paper. To accommodate multiplayer estimation, corresponding improvements are made for time-domain convolution and graph convolution, respectively. Similar to the network structure in [25], four time-domain convolutional modules are constructed, each consisting of convolutional layers such as batch normalization, residual connectivity, ReLU loss function with random deactivation.

There is a graph convolution module between each two time-domain modules, considering the requirement of multiperson estimation, combining the depth information of the target on the basis of time-domain graph convolution, so that the multiperson pose has real spatial distribution and relative position relationship.

4.2. Depth Estimation.

The study in [26] uses a top-down approach to first detect all the key points in the image and then combine the joints belonging to the same target according to the PAF (part affinity field) principle, which can be adapted to multihuman 2D pose estimation of video streams. Before the human pose estimation, there is another branching task-using Mask R-CNN [1] to first presample the person in the video and perform 2D pose estimation in the stationary extended state to obtain a standard skeletal model of each target for 3D pose estimation and correction unit. Then, all human targets in each frame of the video stream are detected, and the character markers with borders are output. The depth estimation of the targets is based on the 2D estimation of the joint point coordinates and the target detection with border regression. For each target segmented by the regression border, the pelvic key point P is selected as the human datum, and the target-to-camera distance (x_p, y_p, z_p) is calculated according to the imaging principle of the pinhole camera according to the method proposed in [11], where z_p denotes the distance to the camera. Lacking support conditions for estimating the distance only from

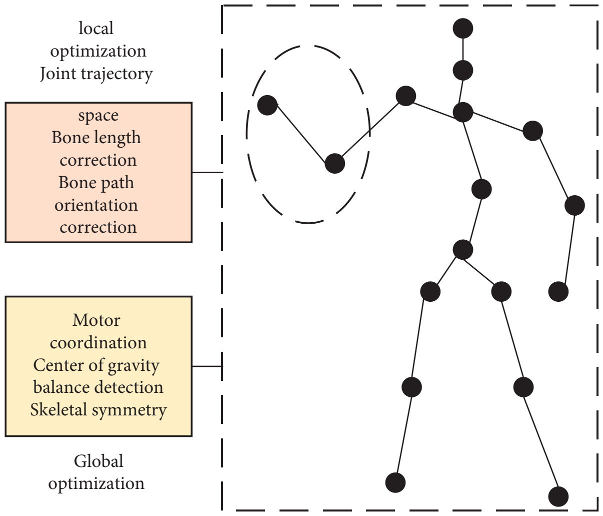

Figure 3: Schematic diagram of the structure of local and global optimization of the human posture.

two-dimensional images, a new measurement scale was designed k as follows:

$$k = \sqrt{\alpha_x \alpha_y \frac{A_{\mathrm{real}}}{A_{\mathrm{img}}}}, \tag{1}$$

where α_x, α_y denotes the focal length divided by the distance factor of X-axis and Y-axis, respectively, and A_{real} and A_{img} denote the area of the human body in real space (in units: mm^2) and image space (in units: pixel^2), respectively. For a given camera parameter, k approximates the absolute depth from the target to the camera by the ratio of the area of the real space to the area of the imaging space. The distance d can then be obtained according to the imaging principle as follows:

$$\begin{aligned} d &= \alpha_x \frac{l_{x,\mathrm{real}}}{l_{x,\mathrm{img}}} \\ &= \alpha_y \frac{l_{y,\mathrm{real}}}{l_{y,\mathrm{img}}}. \end{aligned} \tag{2}$$

Here, $l_{x,\mathrm{real}}$, $l_{y,\mathrm{real}}$, $l_{x,\mathrm{img}}$, and $l_{y,\mathrm{img}}$ denotes the length of the projection of the target on the coordinate axis in the actual space and the image space, respectively. However, this approach has drawbacks; for example, adults and children are in different depth cases, but the sizes are similar on the photos, and the depths output by the depth estimation module are the same. The solution is to involve the depth information in the training of the neural network, which in turn causes difficulties for the existing dataset and requires the addition of depth annotations, and eventually, a new dataset with a small portion of annotated data and a large portion of unannounced data is selected to train the network.

4.3. Contact Calibration. The study in [13] used a method of training neural networks to determine the foot-ground contact, although there is no dedicated public dataset currently available, for which they produced a new dataset. However, on the one hand, producing such a dataset requires a large investment, and on the other hand, the action relationship between the human body and the ground is still difficult to be fully described only by explicit image annotation, while dynamical analysis is an important reference for optimizing the contact problem, so it is proposed to optimize the contact relationship with the help of dynamical analysis in the absence of data and insufficient network training, which also helps to correct the human body imbalance in the next Step [14–16], and a simplification is done as shown in Figure 4. The human body model is simplified into eight parts such as head, torso, left arm, right arm, and left and right leg, and the whole body mass M is initialized to 70 kg, and the mass mi of each part is done according to the general proportion approximate distribution. The contact force between the human body and the ground only considers the internal forces generated by gravity and acceleration of motion. The support force of the ground on the human body comes from the biped, which is represented by $f_r, f_l \in \mathfrak{R}^3$, respectively, and the support force and the combined force of the human body are a pair of interacting forces. Human body dynamics are analyzed to find f_r and f_l.

$$\begin{cases} F = \sum_{i=0}^{J} \left(m_i \ddot{q}_i + m_i g \right), \\ \sum_{i=0}^{J} m_i \ddot{q}_i = \mathrm{M}\ddot{\mathrm{C}}, \\ \mathrm{M}\ddot{\mathrm{C}} = f_r + f_l + Mg, \end{cases} \tag{3}$$

where $q_i \in \mathfrak{R}^3$ are the coordinates of each part of the simplified model of the human body, obtained by averaging the coordinates of the corresponding joints; $\ddot{q}_i$ are their second-order derivatives; and $C, \ddot{C} \in \mathfrak{R}^3$ are the coordinates of the center of mass of the human body and their second-order derivatives.

The initial ground representation needs to be obtained along with the 2D pose estimation. Since it is difficult to label all the training data and, in addition, it is assumed that the plane in which the target is located is a flat area without undulations, only a portion of the data is labeled, which is used to obtain the initial ground on the one hand and to improve the adjustment of the correction unit to the contact state with the ground on the other hand.

Compared to the task of estimating the pose, the acquisition of the ground is relatively simpler and easier. A direct way to determine the foot-ground contact is to check the distance between the coordinates of the foot joint point and the initial ground d. And according to the kinetic analysis, the foot is in contact with the ground when the force of the ground on the body is greater than zero. If the output of contact markers estimated from 2D is wrong, or if the markers are correct, but the output 3D pose is penetrating or vacating with the ground, the pose space position needs to be corrected in combination with the kinetic

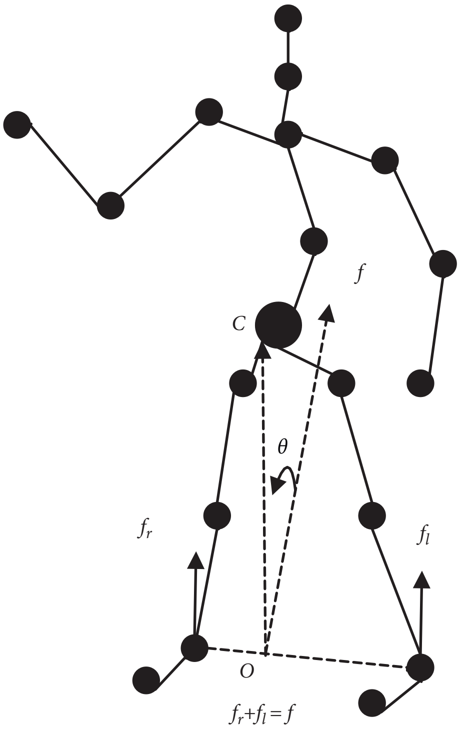

Figure 4: Human body mechanics analysis.

conditions. All joint point coordinates need to be flattened up or down by an offset distance Δd.

4.4. Balance Correction. For the human body balance state, there are static more dynamic. The static situation only needs to consider the effect of ground support and gravity, while the focus is on the dynamic balance. Each joint point of the human body generates momentum, acceleration, rotational moments, and complex forces between each other, and frictional forces on the ground in addition to pressure, which together constitute the equilibrium or imbalance of the human body. In order to maintain the postural balance, the study in [15] proposed controlling the human body balance by momentum calculation.

Motivated by these methods, a geometry-based balance control method is proposed in this paper. Combined with the prediction of the center of mass of the human body, the human body geometry is used to correct the human body balance state. First of all, it is still to determine whether the human body is out of balance, and the prerequisite of the judgment is that the current posture is in contact with the ground, disregarding the action of itself in the vacant state for the time being. The following model is used to analyze the conditions required for human equilibrium. The location of the center of mass C should be near the extension of the combined force *f* of the ground force f_r, f_l on the feet, and O is the intersection of *f* and the ground. The angle θ formed by the center of mass, O and *f* is within the threshold value, and the θ obtained from the imbalanced posture deviates significantly from this threshold value. Based on the measurements, this threshold is taken to be 5°.

$$\cos\theta = \frac{\overrightarrow{\mathrm{OC}}.\overrightarrow{f}}{|\overrightarrow{\mathrm{OC}}|.|\overrightarrow{f}|}. \tag{4}$$

Let the vector $\overrightarrow{k} \perp \overrightarrow{\mathrm{OC}}$ and $\overrightarrow{f}$, except for the joint points of the foot, and rotate around k by the angle θ.

$$\begin{aligned}\overrightarrow{k} &= (x, y, z),\\ \overrightarrow{\mathrm{OC}} &= (a, b, c),\\ \overrightarrow{f} &= (n, p, q).\end{aligned} \tag{5}$$

From the system of chi-squared equations,

$$\begin{cases} ax + by + cz = 0,\\ nx + py + qz = 0.\end{cases} \tag{6}$$

The vector k can be obtained as the vector from the original node to the point O is P, which is rotated around the unit vector u of k to obtain a new node, set as p':

$$p' = p\cos\theta + (u\cdot p)u(1-\cos\theta) + (u\times p)\sin\theta. \tag{7}$$

Loss function: let vector $\widetilde{b}_k$ denote the standard bone length between the nth and mth joints, and b_k be the bone length output from the pose estimation network $b_k = \|p_n - p_m\|$. The loss of bone length prediction is as follows:

$$E_{\mathrm{bone}} = \sum_{k=1}^{b}\left\|b_k - \widetilde{b}_k\right\|. \tag{8}$$

The predicted loss of joint node coordinates is represented by L2 distance; $P_{j-2\,D}$ and $P_{j-3\,D}$ are the predicted 2D/3D joint point coordinates, and $\widetilde{P}_{j-2\,D}$ and $\widetilde{P}_{j-3\,D}$ are the real 2D/3D joint coordinates:

$$E_{\mathrm{point}} = \frac{1}{J}\sum_{j=1}^{J}\left[\left\|P_{j-2\,D} - \widetilde{P}_{j-2\,D}\right\| + \left\|P_{j-3\,D} - \widetilde{P}_{j-3\,D}\right\|\right]. \tag{9}$$

5. Experimentation and Evaluation

5.1. Data Sets and Evaluation Mechanisms. The Human3.6 M dataset is the most commonly used dataset for human pose estimation tasks. It contains 3.6 million single-person video frames captured by a motion capture system in an indoor environment with 11 professional action actors showing 15 daily behaviors (e.g., walking, standing, talking) that can be adapted to single-person pose estimation and camera center coordinate prediction tasks. Based on previous experience, parts 1, 5, 6, 7, and 8 of the dataset are used for training and 9 and 11 are used for testing. The MuPoTS-3D dataset is a multiperson 3D pose estimation dataset that contains more than twenty indoor and outdoor scenes. The realistic 3D motion of each person in the video is derived from a multiview masterless capture that can be adapted to both human-centered and camera-centered coordinate systems. MuCo-3DHP is another dataset for multiperson 3D estimation, which is derived by combining the MPI-INF-3DHP

Table 1: Comparison with other methods using the Human3.6 M dataset under the MPJPE evaluation criteria.

Method	Dir	Dis	Eat	Phon	Pose	Pur	Sit	SitD	Smo	Phot	Wait	Walk
Dario	45.9	48.5	44.3	47.8	51.9	57.8	46.2	45.6	59.8	68.5	50.5	46.5
Cai	46.5	48.8	47.6	50.9	52.9	58.6	58.3	48.3	45.8	59.2	64.4	50.7
Moon	51.5	56.8	51.2	52.2	55.4	47.7	50.9	53.3	68.5	54.7	58.6	60.2
Xu	38.2	44.4	42.8	43.7	47.6	60.3	42.0	45.4	53.2	60.8	46.4	43.5
This article contains 5 frames	43.7	42.3	44.0	45.9	50.1	54.3	54.3	41.8	48.3	54.4	59.2	49.1
This article contains 9 frames	48.3	45.6	49.4	46.7	52.9	57.1	42.4	50.4	56.2	62.3	56.3	62.5

Table 2: Quantification of prediction accuracy on the MuPoTS-3D dataset.

Method	MP TPE	P-MPTPE
Basic network	55.4	47.5
Posture correction	51.3	43.7
Contact correction	52.1	43.6
Balance correction	52.4	42.9
Whole network	51.4	42.3

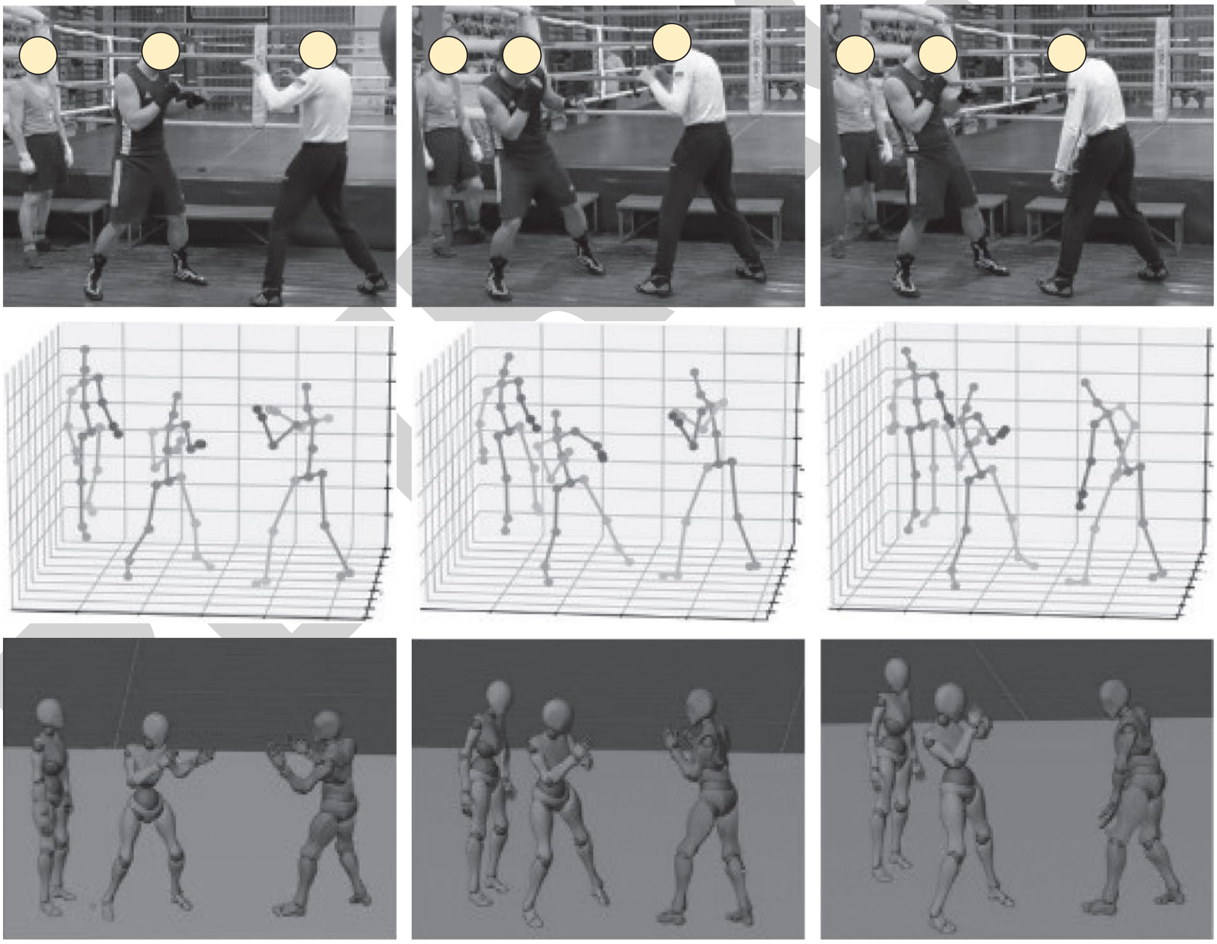

Figure 5: Motion capture visuals on the COCO2017 dataset.

3D single-person dataset. The MuPoTS-3D dataset is used for testing, and MuCo-3DHP is used for training.

Evaluation protocols: there are two evaluation protocols that are widely used: the first one calculates the mean error of predicted pose and true pose joint point coordinates (MPJPE), and the second one is the mean error after alignment [11] and this mechanism is called PA MPJPE [27].

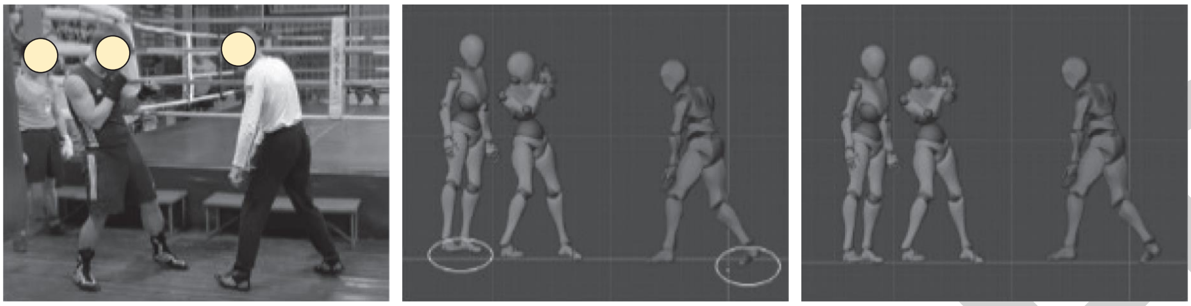

Figure 6: Contact correction effect.

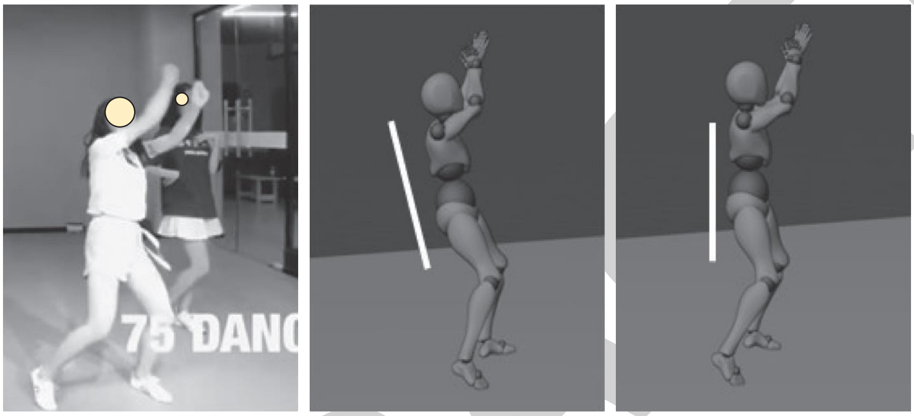

Figure 7: Balance correction effect.

5.2. Realization Process. The Mask R-CNN exposed models used in this paper are pretrained on the coded dataset for target detection and age regression. For time-domain graph convolutional networks, their main parts are initialized with the publicly available ResNet-50, which is pretrained on the ImageNet dataset [15] is performed in training using a 256 × 256 size input image. The training of the pose estimation network is divided into two phases: the first phase uses the human 3.6 M dataset, and the 5th and 64th frames of each video are used for testing, using the MPII dataset in addition to the Human3.6 M dataset, with half of each of the two datasets in each training set. The MuCo-3DHP and MuPoTS-3D datasets were used in the second stage, and half of the data in each batch was COCO data in order to enhance the dataset. The experiments were conducted using four NVIDIA 1080 Ti GPUs for 20 rounds of training in the time-domain graph convolution network.

Experimental results on publicly available datasets show that the pose accuracy of the model output in this paper reaches the level of single-person video estimation, and the accuracy and smoothness of the output are related to the number of targets and frame rate. The data in Table 1 show that the accuracy of the monocular multiplayer video estimation model proposed in this paper is comparable to that of current single-person estimation models.

5.3. Calibration Module Assessment. Ablation results of the three correction units were compared on multiple evaluation protocols, using two evaluation protocols [28]. The data show that pose correction makes the largest contribution to accuracy improvement, with limited contributions from leveling correction and contact correction, but it is critical to the practical application of motion capture. The effect of each component on the final results is shown in Table 2, the multiplayer motion capture results are shown in Figure 5, and the correction units are shown in Figures 6 and 7.

6. Conclusions

The paper mainly discusses the application of modern motion capture technology in the production of film and television animation. With the continuous development and improvement of motion capture technology, motion capture technology will surely get more and more important applications in our daily lives. The emergence of optical motion capture systems has greatly reduced the cost of filming and animation production and has made film and animation pictures realistic. The motion capture model in this article can capture the actions of multiple people in a nonrestricted environment. After the human body pose is corrected, it basically conforms to the real human motion. It works well in three-dimensional virtual characters. It does not require professional shooting equipment and markings. In the case of other auxiliary equipment, materials can be obtained in daily life, and the simplified design of the motion capture system is basically realized. We will focus on the realization of three-dimensional estimation, especially the multiperson goal. Time-domain convolution makes full use of the key information at different time points of the video stream to infer 3D pose. Regarding the connection relationship of key points of the human body as a graph structure is a prerequisite for realizing graph convolution, and then, the 3D relationship of bones can be extracted from the global and local.

In the future, we need to pay attention to the preprocessing of data, especially the selection of data features, and the optimization of network models.

Data Availability

The dataset used in this paper is available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflicts of interest regarding this work.

References

[1] Z. Manyu, "Application of performance motion capture technology in film and television performance animation," *Applied Mechanics and Materials*, vol. 347-350, pp. 2781–2784, 2013.

[2] T. Lauthelier and M. Neveu, "Facial animation by reverse morphing on a sequence of real images: application to film and video production," *Annales des Telecommunications*, vol. 55, no. 3/4, pp. 143–148, 2000.

[3] R. Zeng, "Research on the application of computer digital animation technology in film and television," *Journal of Physics: Conference Series*, vol. 1915, no. 3, Article ID 032047, 2021.

[4] S.-F. Xie, "Study on the creative development of China cartoon & animation - focusing on changes in a production environment with supporting the policy of the state administration of press, publication, radio, film and television of China," *Cartoon and Animation Studies*, vol. 35, pp. 209–226, 2014.

[5] B. Feng, W. Li, J. Shi, and Z. Li, "Framework study on the three-dimensional long-distance running sport training based on the markerless monocular videos," *Revista de la Facultad de Ingenieria*, vol. 32, no. 11, pp. 454–461, 2017.

[6] E. Akpinar, "The use of interactive computer animations based on POE as a presentation tool in primary science teaching[J]," *Journal of Science Education and Technology*, vol. 23, no. 4, pp. 527–537, 2014.

[7] W. Jian, "The analysis of the virtuality of film and television animation art," in *Proceedings of the 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA)*, pp. 119–121, IEEE, Ottawa, ON, September 2014.

[8] C. Cao, Y. Tang, D. Huang, W. Gan, and C. Zhang, "IIBE: an improved identity-based encryption algorithm for wsn security," *Security and Communication Networks*, vol. 2021, Article ID 8527068, 8 pages, 2021.

[9] Y. Guo, B. Li, and K. Fan, "Support vector machine wavelet blind equalization algorithm based on improved genetic algorithm," *Advances in Intelligent and Soft Computing*, vol. 149, pp. 161–166, 2012.

[10] J. X. Huang, W. U. Wei-Long, and C. J. Long, "Study of moving object detection in video and its application based on OpenCV," *Computer Technology and Development*, vol. 135, no. 3, pp. 556–561, 2014.

[11] Y. Tong, W. Cao, Q. Sun, and D. Chen, "The use of deep learning and VR technology in film and television production from the perspective of audience psychology," *Frontiers in Psychology*, vol. 12, p. 501, 2021.

[12] C. Mo, K. Hu, S. Mei, Z. Chen, and Z. Wang, "Keyframe extraction from motion capture sequences with graph based deep reinforcement learning," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5194–5202, Chengdu, China, 2021, October.

[13] D. Zhou, X. Feng, P. Yi et al., "3D human motion synthesis based on convolutional neural network," *IEEE Access*, vol. 7, Article ID 66335, 2019.

[14] J. Lin, J. Cui, G. Shi, and D. Liu, "CG animation creator: auto-rendering of motion stick figure based on conditional adversarial learning," in *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 341–352, Springer, Xian, China, November 2019.

[15] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: a survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, Article ID 102897, 2020.

[16] M. Y. Zhang, "Application of performance motion capture technology in film and television performance animation," in *Applied Mechanics and Materials*vol. 347-350, , pp. 2781–2784, Trans Tech Publications Ltd, 2013.

[17] R. Zeng, "Research on the application of computer digital animation technology in film and television," in *Journal of Physics: Conference Series*vol. 1915, no. 3, IOP Publishing, Article ID 032047, 2021.

[18] N. Du and C. Yu, "Research on special effects of film and television movies based on computer virtual production VR technology," in *Proceedings of the 2020 International Conference on Computers, Information Processing and Advanced Education*, pp. 115–120, Ottawa ON, Canada, October 2020.

[19] N. Du and C. Yu, "Application and research of VR virtual technology in film and television art," in *Proceedings of the 2020 International Conference on Computers, Information Processing and Advanced Education*, pp. 108–114, Ottawa ON, Canada, October 2020.

[20] X. I. E. Tao, C. Zhang, and Y. Xu, "Collaborative parameter update based on average variance reduction of historical gradients[J]," *Journal of Electronics and Information Technology*, vol. 43, no. 4, pp. 956–964, 2021.

[21] R. Ge, H.-Y. Chen, T.-C. Hsiao, Y.-T. Chang, and Z.-Y. Wu, "Virtual reality, augmented reality and mixed reality on the marketing of film and television creation industry," in *Proceedings of the International Conference on 5G for Future Wireless Networks*, pp. 464–468, Springer, Tianjin, P. R. China, August 2020.

[22] L. Wang, C. Zhang, Q. Chen et al., "A communication strategy of proactive nodes based on loop theorem in wireless sensor networks," in *Proceedings of the 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP)*, pp. 160–167, IEEE, Wanzhou, China, November 2018.

[23] H. Li, D. Zeng, L. Chen, Q. Chen, M. Wang, and C. Zhang, "Immune multipath reliable transmission with fault tolerance in wireless sensor networks," in *Proceedings of the International Conference on Bio-Inspired Computing: Theories and Applications*, pp. 513–517, Springer, Xi'an, China, October 2016.

[24] J. Zeng, X. He, Y. Hu, Y. Zhang, H. Yang, and S. Zhou, "Research status of data application based on optical motion capture technology," in *Proceedings of the 2021 2nd*

International Conference on Artificial Intelligence and Information Systems, pp. 1–8, Chongqing China, May 2021.

[25] W. Carroll, A. Turner, P. Talegaonkar et al., "Closing the wearable gap-Part IX: validation of an improved ankle motion capture wearable," *IEEE Access*, vol. 9, Article ID 114036, 2021.

[26] M. N. H. Yunus, M. H. Jaafar, A. S. A. Mohamed, N. Z. Azraai, and M. S. Hossain, "Implementation of kinetic and kinematic variables in ergonomic risk assessment using motion capture simulation: a review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 16, p. 8342, 2021.

[27] Z. Dou, "Research on virtual simulation of basketball technology 3D animation based on FPGA and motion capture system," *Microprocessors and Microsystems*, vol. 81, Article ID 103679, 2021.

[28] Z. Lin, "Research on film animation design based on inertial motion capture algorithm," *Soft Computing*, vol. 25, no. 18, Article ID 12505, 2021.