Hindawi
Advances in Mathematical Physics
Volume 2021, Article ID 3772358, 11 pages
https://doi.org/10.1155/2021/3772358

Research Article

Three-Dimensional Diffusion Model in Sports Dance Video Human Skeleton Detection and Extraction

Zhi Li

School of Media and Art Design, Guilin University of Aerospace Technology, Guilin 541004, China

Correspondence should be addressed to Zhi Li; lizhiys@guat.edu.cn

Received 26 July 2021; Revised 11 August 2021; Accepted 14 August 2021; Published 7 September 2021

Academic Editor: Miaochao Chen

The research in this paper mainly includes as follows: for the principle of action recognition based on the 3D diffusion model convolutional neural network, the whole detection process is carried out from fine to coarse using a bottom-up approach; for the human skeleton detection accuracy, a multibranch multistage cascaded CNN structure is proposed, and this network structure enables the model to learn the relationship between the joints of the human body from the original image and effectively predict the occluded parts, allowing simultaneous prediction of skeleton point positions and skeleton point association information on the one hand, and refinement of the detection results in an iterative manner on the other. For the combination problem of discrete skeleton points, it is proposed to take the limb parts formed between skeleton points as information carriers, construct the skeleton point association information model using vector field, and consider it as a feature, to obtain the relationship between different skeleton points by using the detection method. It is pointed out that the reorganization problem of discrete skeleton points in multiperson scenes is an NP-Hard problem, which can be simplified by decomposing it into a set of subproblems of bipartite graph matching, thus proposing a matching algorithm for discrete skeleton points and optimizing it for the skeleton dislocation and algorithm problems of human occlusion. Compared with traditional two-dimensional images, audio, video, and other multimedia data, the 3D diffusion model data describe the 3D geometric morphological information of the target scene and are not affected by lighting changes, rotation, and scale transformation of the target and thus can describe the realistic scene more comprehensively and realistically. With the continuous updating of diffusion model acquisition equipment, the rapid development of 3D reconstruction technology, and the continuous enhancement of computing power, the research on the application of 3D diffusion model in the detection and extraction of a human skeleton in sports dance videos has become a hot direction in the field of computer vision and computer graphics. Among them, the feature detection description and model alignment of 3D nonrigid models are a fundamental problem with very important research value and significance and challenging at the same time, which has received wide attention from the academic community.

1. Introduction

With the rapid development of 3D sensors, such as structured light coding and LiDAR, the acquisition of 3D diffusion model data has become increasingly convenient and fast in recent years. Diffusion model data is mathematically abstractly described as a collection of three-dimensional coordinates of points, which is essentially a discrete sampling of geometric information of the external world in a specific coordinate system. Compared with traditional 2D images, 3D diffusion model data have the following significant advantages.

1.1. Describe the 3D Geometric Morphological Information of the Target. Traditional 2D images describe the appearance of the external scene, losing 3D spatial information. Diffusion model data describe the 3D geometry of the target surface and thus can more directly inform computer vision tasks such as feature extraction and matching.

1.2. Unaffected by Changes in External Light. Most of the common 3D imaging sensors use active imaging, such as structured light sensors and LiDAR. Therefore, the change of light in the external world does not affect the acquisition of diffusion model data.

1.3. Less Influenced by Imaging Distance. The traditional 2D image imaging process is susceptible to changes in imaging distance, resulting in changes in the scale of the imaged target. The diffusion model data is a discrete sampling of the 3D geometry of the target surface in the external scene, and the imaging distance does not change the scale of the imaged target, but only affects the accuracy and resolution of the acquired data, and thus is more suitable for computer vision tasks.

In recent years, along with the rapid development of 3D reconstruction technology, it has become increasingly convenient to obtain 3D models through 3D data [1]. Since there are many nonrigid objects in the real world, the study of 3D nonrigid models is receiving widespread attention and has become a research hotspot in the fields of computer vision and computer graphics.

The study of human skeleton detection in sports dance video images has been a very popular research direction in image processing and computer vision [2]. The human skeleton information can greatly help people analyze the behavior of the target human body in pictures or videos and lay the foundation for further processing of images and videos [3]. The human skeleton detection algorithm divides the human skeleton into multiple joints, such as head, shoulder, and wrist and then analyzes the position, direction, and movement of each joint to obtain the human skeleton information. The human skeleton is drawn to further analyze the posture and behavior of the human body to obtain the activity and motion information of the human body in the image [4].

Applications related to human posture estimation are based on the premise of obtaining a clear and accurate human skeleton in the image, and inaccurate skeleton extraction will lead to incorrect analysis of human behavior and movements, with incalculable consequences [5]. For example, in the field of sports dance, inaccurate skeleton extraction may lead to incorrect analysis of action, which may even endanger the lives of athletes or performers in serious cases. Therefore, it is of great importance to improve the accuracy of human skeleton detection. In recent years, the rapid development of the hardware field makes the computer's computing power increase, increasingly excellent human skeleton detection algorithms emerge, and the human skeleton detection accuracy is continuously improved. As the basis of human pose recognition, human skeleton detection technology will play an increasingly important role in increased fields.

2. Related Work

Since the 1970s, the study of geometric morphological information of target 3D diffusion models has been receiving attention, and a series of results have been achieved in the 1980s and 1990s. The detection of saliency regions for 3D diffusion geometry models is a complex problem, especially for 3D diffusion models with isometric transformations [6]. In recent years, the problem has been intensively investigated in the fields of computer vision and computer graphics. The literature [7] first started to address the problem of 3D deformable model region detection by describing it abstractly as finding the most stable component on the model. To have invariance to isometric transformations, the method uses diffusion geometry to derive weighting functions and proposes two representations of mesh surfaces, namely, mesh vertex-based and edge-weighted graph structures, respectively [8]. Experimental results are realized that the edge-weighted graph structure-based representation is more general than the vertex-weighted graph and exhibits superior performance [9]. The algorithmic framework has been extended to handle shapes with volumes. Inspired by cognitive theory, the literature [10] considers saliency regions as "key components" on the model and considers that they contain rich and distinguishable local features. According to this theory, saliency regions correspond to parts of the model with high protrusions and can be detected by a clustering process in geodesic space. However, this method is an incomplete decomposition of the model and many regions of saliency are not detected [11]. The method based on diffusion geometry has achieved remarkable success in the analysis of 3D nonrigid models due to the reflection of the model's intrinsic properties [12]. The literature [13] combined the eigenfunctions of the Laplace Beltrami operator with homology consistency theory to generate a hierarchical segmentation method for the model. The literature [14] first calculates the global signature of each point on the model, then maps the model into its eigenspace, and finally uses a clustering algorithm in that space to achieve the segmentation of the model. All the above algorithms use the eigenfunctions of the Laplace Beltrami operator for model segmentation; however, the eigenfunctions are prone to problems such as significant change or eigenvector switching, especially when the differences between the corresponding eigenvalues are small [15]. The literature [16] introduces the idea of consensus clustering into this domain to achieve stable segmentation. First, multiple clusters are computed in the global point signature space to generate a heterogeneous set of model partitions. The literature [17] argues that a stable model segmentation can be obtained by extracting statistical information from these segmentations. This method has the best current results in the case of model data receiving various disturbances.

Human skeleton detection in images can be divided into two directions: 2D human skeleton detection and 3D human skeleton detection. 3D human skeleton detection is the process of obtaining the 3D shape or coordinates of human skeleton points by analyzing the images obtained from 3D cameras such as Kinect.

3. Application of THE Three-Dimensional Diffusion Model in the Detection and Extraction of a Human Skeleton in Sports Dance Videos

3.1. Principle of Action Recognition Based on THE 3D Diffusion Model Convolutional Neural Network. The 3D diffusion model neural network is one of the first deep learning methods to achieve great success in the fields of image

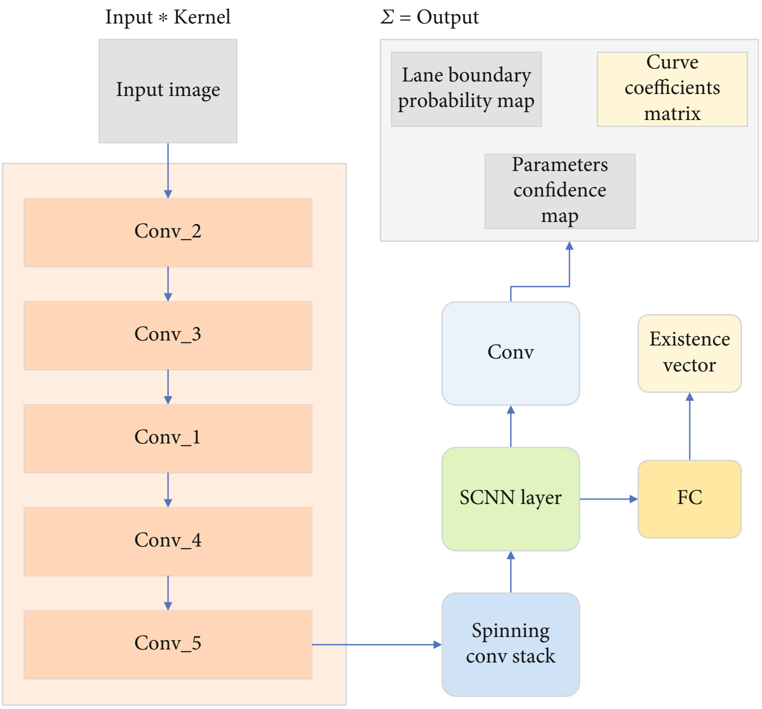

Figure 1: The traditional two-dimensional convolution process.

analysis, target detection, and so on. It applies trainable filters (trained by backpropagation algorithm), local domain pooling operations (to prevent overfitting), etc. in the original input to extract gradually complex and highly abstract input features, and the network model can achieve very good discriminative effects through long training with a large amount of data [18]. And it also has lighting, background, pose extraction invariance, and other characteristics, which are very popular.

As an exemplary end-to-end network model, convolutional neural networks can produce effects directly on the original input, which makes the traditional manual extraction of features outdated. However, currently, such convolutional neural networks are still only heavily used in fields such as input recognition of 2D images, and Figure 1 illustrates the traditional 2D convolution process. To make greater use of its power, some groups extended it to the 3D domain, generating a new 3D diffusion model and applying it to the subject of human action recognition, producing very good results. Its main feature is that it is not only able to extract features in space but also combines feature extraction in the temporal dimension, using 3D convolution to capture human motion information in consecutive frames.

The main difference between 3D convolution and 2D convolution is the difference between the perceptual field and the convolution kernel. 3D convolution takes the same frame parts of multiple consecutive frames and forms a special cube and then performs a convolution operation in the cube using a 3D convolution kernel. This means that in a multilayer convolution operation, the input feature map of the next layer is related to the multiple video frames that form the cube in the previous layer to capture information in the temporal dimension of the video frames. As shown in Figure 2, its input feature map consists of the same local images from three adjacent video frames in the upper layer together. Like 2D convolution, 3D convolution also requires several different convolution kernels to extract different feature information in the spatiotemporal features. As the number of convolution layers increases, we can extract more types of high-level abstract features from multiple combinations of primary feature maps.

Suppose $[h_1, \mathrm{b}_1, \cdots, h_i, \mathrm{b}_i]$ is a given training sample with labels, a total of m, h, and b_x are the output value of the network, i.e., one prediction of the model for the input sample x . The purpose of training the neural network is to make the predicted value h_w, b_x as close as possible to the true value y. The error between the two can be represented by the loss function. For a single sample (x, y), the variance loss function can be expressed as

$$J(w, b) = \frac{1}{2} \lim_{m \longrightarrow \infty} \sum_{i=1}^{m} \left[h_{w,b}(x) - y_i\right]^2. \tag{1}$$

For the entire training sample set, the cost function is

$$J(W, b) = \lim_{m \longrightarrow \infty} \frac{1}{m} \sum_{i=1}^{m} J(w, b_i) + \frac{\lambda + 1}{2} \sum \left(w_{ij}\right)^2, \tag{2}$$

where i is the number of network layers, and j is the number of nodes in the jth layer. The loss function includes both mean squared deviation and weight decay, and the purpose of introducing the weight decay term is to prevent overfitting during the training process. The loss function represents the difference between the true value and the predicted value, and the smaller the difference represents,

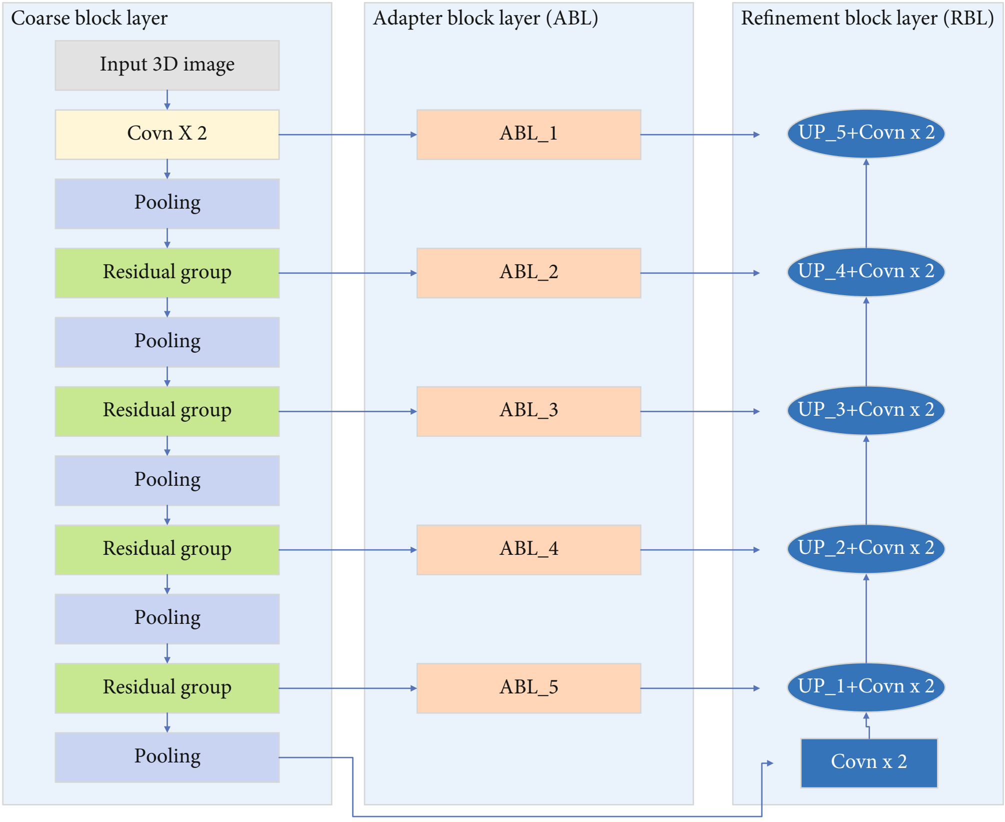

Figure 2: The traditional 3D convolution process.

the more accurate prediction of the network. The final training goal of the neural network is to solve the weight parameter W and the bias parameter b such that the loss function $J(w, b)$ is minimized. The gradient descent algorithm (GDA) is a common algorithm for finding the optimal solution of parameters in neural networks, and its parameter iterative update process is

$$W_{ij}{}^{l} = -\alpha \sum \frac{\sqrt{J(w, b)}}{\partial w}, \tag{3}$$

$$b_{ij}{}^{l} = b_{ji}{}^{l} - \alpha \sum \frac{\partial \sqrt{J(w, b)}}{\partial b^{l}}, \tag{4}$$

where α is the learning rate, which indicates the magnitude of each update of the parameters. The core of the gradient descent algorithm is to derive the network parameters and spread the gradient upward layer by layer. Taking a single sample (x, y) as an example, its loss function $J(w, b\,; x^{i}, y^{i})$ is biased concerning W and b as

$$\frac{\partial}{\partial W_{ij}{}^{l}} J(\mathrm{w}, \mathrm{b}) = \frac{1}{b} \lim_{n \longrightarrow \infty} \sum_{i=1}^{m} J\left(w, b\,; x^{i}, y^{i}\right) + (\lambda - 1) W_{ij}{}^{l}. \tag{5}$$

δ_{l} is obtained by calculating a weighted average of the errors of the nodes at the $(\mathrm{l}+1)$st level, it represents the number of nodes at the lth level. From Equations (6) and 7, the partial derivatives of the loss function $J(w, b)$ concerning each parameter are

$$\nabla_{w^{t}} J(w, b\,; x, y) = \alpha^{l} \delta^{l+1} \frac{\partial J(w, b)}{\partial w_{ij}{}^{l}}, \tag{6}$$

$$\nabla_{b^{t}} J(w, b\,; x, y) = \delta^{l+1} \frac{\partial J(w, b)}{\partial b_{ij}{}^{l}} + \lambda \alpha_{ij}{}^{l}. \tag{7}$$

In the traditional method of extracting a 3D skeleton from a single depth map, the general steps are firstly extracting the body features, secondly classifying the body features by different parts, and finally locating the joint points to generate the 3D skeleton of the human body. Unlike 2D skeleton extraction, the depth map is better able to deal with problems such as body self-occlusion because it contains depth information, but it is still more challenging to accurately predict the fixed positions of different joint points in 3D space. Extracting specific image features from the depth map is an important part of the whole process. To be able to shorten the computing time as much as possible, the extracted image dimension should not be too large, and the features should have strong representational characteristics that can well distinguish different classes of samples. Common depth map features somewhat features include SIFT, SURF, gradient features such as Canny operator, and gradient histogram with direction, etc. In the literature, the authors creatively combined point features and gradient features, and this method cannot only reflect the surrounding information of the feature pixel points without losing the depth of feature information.

The specific operation is similar to CNN convolution, with a pixel point x in the depth map as the center, $f_\phi(p, x)$, $f_\phi(q, y)$, and x, y is the depth value at that point, then there are 8-pixel points adjacent to it, representing 8 different directions, with the horizontal to the right direction as the reference, λ_{θ_1} denotes the angle between any direction vector and its, for each vector pair φ, and the feature calculation formulas are

$$f_\phi(p, x) = d_p\left(x - \frac{\lambda_{\theta_1}}{d_p(x)}\right) + d_p\left(x - \frac{\lambda_{\theta_2}}{d_p(x)}\right), \tag{8}$$

$$f_\phi(q, y) = d_q\left(y - \frac{\lambda_{\theta_1}}{d_q(y)}\right) + d_q\left(y - \frac{\lambda_{\theta_2}}{d_q(y)}\right) + \lambda. \tag{9}$$

This feature is not only computationally small but also has displacement-invariant spatial characteristics, which can be used to extract features from the images in the training set.

To make the network order invariant for unordered 3D data, Point Net uses simple symmetric functions to obtain global features in 3D. For the asymmetric function, the output value does not change with the order of the input variables, e.g., the function $g(x, y, z) = x + y + z$ is symmetric, and the final function value is the same regardless of the order of the independent variables. The formula for calculating the 3D global features in Point Net is

$$f(\{x_1, x_2, \cdots, x_n\}) = g(h(x_1) + h(x_2) + \cdots + h(x_n)). \tag{10}$$

The commonly used activation functions are the Sigmod function, ReLU function, Tanh function, etc. The activation functions are mainly used to introduce nonlinear factors into the neural network; otherwise, the neural network is just a linear combination of inputs [19]. In this paper, we mainly use the Sigmod activation function and the ReLU activation function. The expression of the Sigmod function is shown in Equation (11), which can map the input values between 0 and 1 and can be used in the attention mechanism for weight assignment, and the Sigmod function is derivable and can be optimized with the gradient backward iteration algorithm. Its derivative is shown in Equation (12).

$$f(x) = \min\,(x, 0), \tag{11}$$

$$f(x) = \min\,(x, 0)\begin{cases}1\\2\end{cases}. \tag{12}$$

However, from Equations (8)-(12), the derivative becomes smaller as x becomes larger and converges to 0 as x tends to infinity. Therefore, when training with the backpropagation algorithm, the gradient becomes smaller and smaller as the network deepens, causing the gradient to vanish.

The boundary area of the 3D CAD model described here is the triangular mesh. The inertia tensor of the object is

$$A = \left(a_{ij}\right)_{3\times3}. \tag{13}$$

The equation for the center of mass of an object is

$$(x, y, z) = \lim_{k\longrightarrow\infty}\frac{2}{k+1}\left(\sum_{i=1}^{k}\lambda(x_i + y_i + z_i),\right. \tag{14}$$

$$(\alpha, \beta, \gamma) = \lim_{k\longrightarrow\infty}\frac{2}{k+1}\sum_{i=1}^{k}\left(\frac{\alpha_i + \beta_i + \gamma_i}{3}\right). \tag{15}$$

The rigid transformation of the 3D geometric model can be represented by using a matrix, while the no-rigid transformation of the model can only be described by point matching; so, the 3D nonrigid model alignment is essentially a combinatorial optimization problem with high complexity. Meanwhile, in real scenarios, 3D geometric model data may be subject to various disturbances, such as isometric transformations, holes, small holes, scale transformations, local scale transformations, resampling, noise, scattered grain noise, and topological transformations, which requires the alignment algorithm to have strong robustness.

3.2. Human Skeleton Detection Based on THE 3D Diffusion Model Algorithm. Usually, if we can infer 3D skeleton information from 2D skeleton information, it must be driven by high dimensional knowledge, e.g., anthropometric, kinematic, and kinetic constraints. Some groups transformed this problem into a regression problem by learning a regression mapping model for 3D poses by integrating spatiotemporal integration features in sequential pictures. Many authors have converted this problem into a typical constrained optimization problem to minimize the main error in mapping an unknown 3D pose to a 2D pose at an unknown observation angle. This optimization problem is subject to the corresponding application constraints and sometimes requires the assumption that the 3D pose has a better optimization state in the lower dimensional subspace, but this optimization-based approach can be sensitive to initialization and local minima and often require expensive constraint solvers [20]. With the rapid development of deep learning, we are more pleased to find that many algorithmic effects change qualitatively when the amount of data reaches a certain level. Inspired by this data-driven architecture, we can improve the constraints of previous algorithms by changing them to a simple nonparametric encoding of high-level constraints. This approach is implemented thanks to the availability of a large sample of 3D skeleton information dataset, and the whole algorithm flow is as follows: given a 3D pose library, we generate many 2D projections (from a virtual camera view) and build a paired (2D, 3D) human pose library. The process of using a virtual camera to obtain the 2D projections, which draws on the 3D skeleton extraction method based on a monocular camera, is

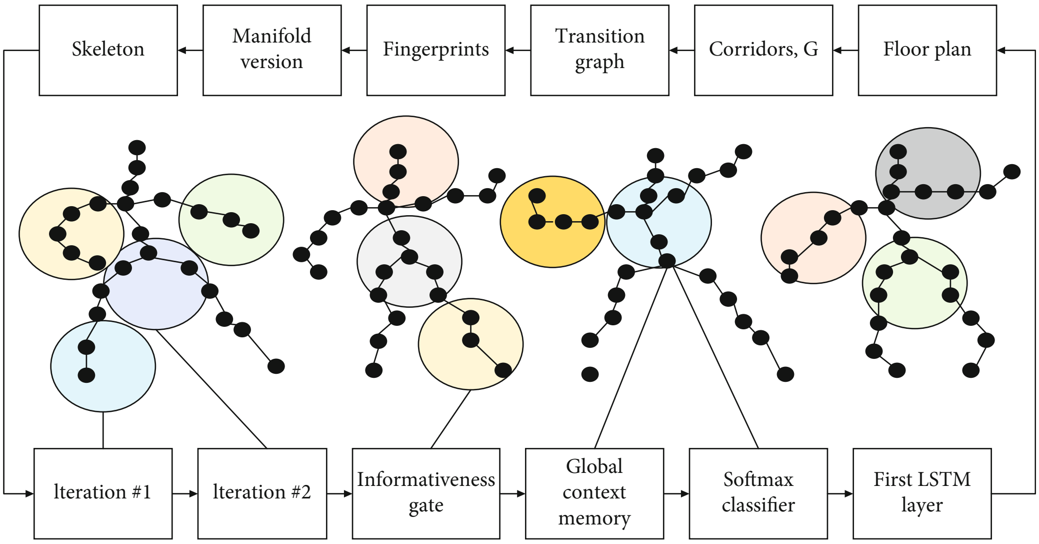

FIGURE 3: The matching-based 3D skeleton extraction process.

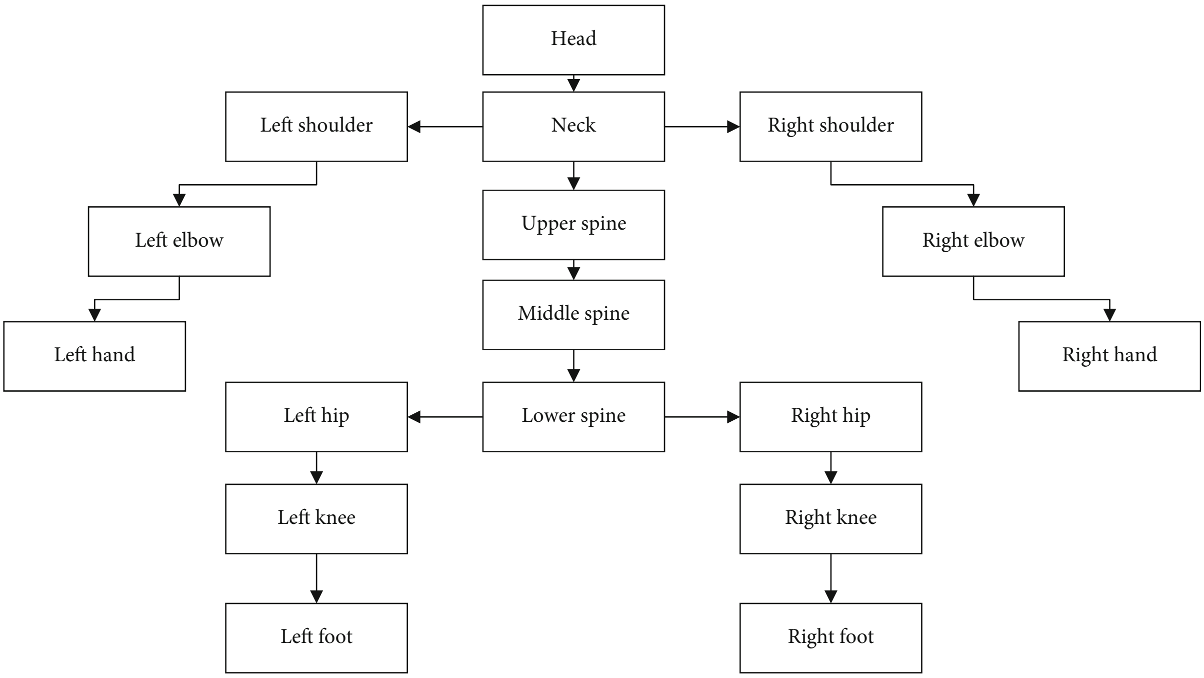

FIGURE 4: Joint points after skeleton identification.

achieved by linking the 3D skeleton, the 2D skeleton to the original picture, and the depth information to the sparse representation. With these paired (2D, 3D) data, and the 2D skeleton extraction results obtained from the pictures by the general 2D pose evaluation algorithm, using the corresponding matching algorithm, we can invert the paired items that are closest to the 2D samples from the paired database and obtain their corresponding 3D skeleton data information. Therefore, the whole algorithm process can be briefly summarized into two parts: the first part extracts the corresponding 2D skeleton from the image by deep learning method, and the second part finds the corresponding 3D skeleton data information from the 3D action database by using the matching algorithm. The specific process is shown in Figure 3.

Thus, we can shorten the time of the whole matching process by reducing the matching range at each matching, thus improving the efficiency of the whole matching algorithm while still considering the accuracy rate. At the same time, as the human body is in motion, there is a faster movement, when the direct use of the above algorithm will lead to a larger recognition error. Considerin, to ensure the reliability of the algorithm operation, every L frame, we research the whole sample space; that is, we can get a more accurate 3D pose. The 14 specific nodes after recognition are shown in Figure 4.

The bottom-up pipeline for 3D skeleton point detection also consists of two parts: skeleton point detection and skeleton point clustering. It is a “fine to coarse” process, i.e., all skeleton points in the image are first detected, and then the detected skeleton points are clustered by some related strategies to form individual by individual [21]. The difference is that the skeleton point detection here needs to detect all skeleton points of all categories in the image at once.

The bottom-up detection process is straightforward. It detects the skeleton points only once for the whole image; so, the running time is independent of the number of people in the image. However, this method has a discrete point aggregation process, which requires finding which person all skeleton points belong to. This is an NP-Hard problem for solving integer linear programming on a fully connected graph, with an average processing time of several hours (e.g., Deepcut, which is a bottom-up approach, takes 50,000 seconds to process a single image). Compared with the top-down method, the bottom-up method is not affected by the number of bodies in the image, has faster detection speed and stronger robustness, and its detection results are superior to the top-down method if the relationship between skeleton points can be effectively constructed, which is the proposed detection method in this paper.

4. Experiments and Result Analysis

The experimental platform used in this paper is TensorFlow 1.6, and the training is performed with a single graphics card, NVIDIA Tesla V100 (16G). The first stage subnetwork uses a batch size of 16 and converges after 40 epochs for the UBC3V dataset and 6000 epochs for the dataset provided in this paper. The second stage subnetwork uses a batch size of 8 and converges after 30 epochs for the UBC3V dataset and 4000 epochs for the dataset provided in this paper. The initial learning rate is set to 0.001, and the learning rate decreases to 90% of the original rate after every 5000 iterations. In this paper, data augmentation is adopted during training, 2048 points are randomly sampled from the human 3D model as the input of the network before each training, and the 3D data are randomly rotated before the input, where the rotation angle around the y-axis (vertical axis) is randomly chosen from [-180°,180°], and the rotation angles around the x-axis and y-axis are randomly chosen from [-20°,20°]. For the UBC3V dataset, this paper conducts training, validation, and testing according to the training set, validation set, and test set divided in advance by this dataset. For the small-scale dataset provided in this paper, a five-fold crossvalidation method is used in this paper to evaluate the algorithm this paper. In the test, the human 3D model containing 8192 points is randomly divided into 4 human 3D models containing 2048 points and then input to the first stage subnetwork for disambiguation to obtain a nonambiguous human 3D model consisting of about 4000 points, and it is randomly divided into 2 nonambiguous human 3D structure groups containing 2048 points (by random repetitive sampling to ensure the number of points). The node prediction 3D model is then input to the second-stage subnetwork to obtain the node prediction 3D model. Finally, the 3D human skeleton is obtained by filtering and aggregating the predicted 3D models.

The method in this paper takes the human body 3D model as input; although it cannot process the depth image directly, it can convert the depth image into point cloud before processing; so, this section compares the method in this paper with the traditional point cloud curve skeleton extraction method and the human skeleton estimation method based on the traditional method, respectively.

From the experimental results, the human skeleton obtained by the LBC, L1, and MDCS algorithms is generally correct for the human point cloud model with separate limbs, but there are also a few missing branches, redundancies, and broken skeletons. For the human pose with close body parts, body contact, or limb crossing, the skeletons extracted by LBC, L1, and MDCS algorithms contain more errors. This is because these traditional point cloud skeleton extraction methods are not able to perceive semantic information, and when body parts are close or in contact, these methods may ignore some human body structures or perceive wrong structures; for example, the human arm naturally drops down close to the body, when the traditional methods cannot perceive the arm branches, resulting in the missing arm in the extracted skeleton. In contrast, the deep learning-based method proposed in this paper can perceive the semantic information of body parts, and the corresponding joint points are predicted for each body part; so, the extracted skeleton is more accurate, where the semantic information of joint points is represented by different colors. The method in this paper can obtain a more accurate 3D human skeleton for both simple and more complex human body poses.

In addition, to verify the robustness of this method to missing data, different algorithms are used to extract the human skeleton on the point cloud data with missing points, and the skeleton extraction results of different traditional point cloud curve skeleton extraction algorithms and this algorithm on the human point cloud with missing nonarticular parts, partially missing joint parts and completely missing joint parts. It can be seen that the skeleton extracted by the three traditional algorithms will have the corresponding branch missing or branch offset when the nonjoint part is missing or the joint part is partially missing, but the method in this paper can still get more accurate results because the disambiguation strategy in the first stage of this paper has removed the points far from the joint part, which makes the data used in the second stage of subnetwork training. This makes the data used in the second stage subnetwork training the human point cloud with a large number of missing points in nonjoint areas. Therefore, if the missing points occur in nonjoint areas, the impact on the method in this paper is minimal. This is because there are multiple predicted values for each joint point, and even if the surface points of a joint point are missing on one side, the surface points on the other side will still shrink to the corresponding joint point to obtain the predicted value for that joint point. However, when all the points of a joint part are missing, the skeleton extracted by the method in this paper will also show branch missing.

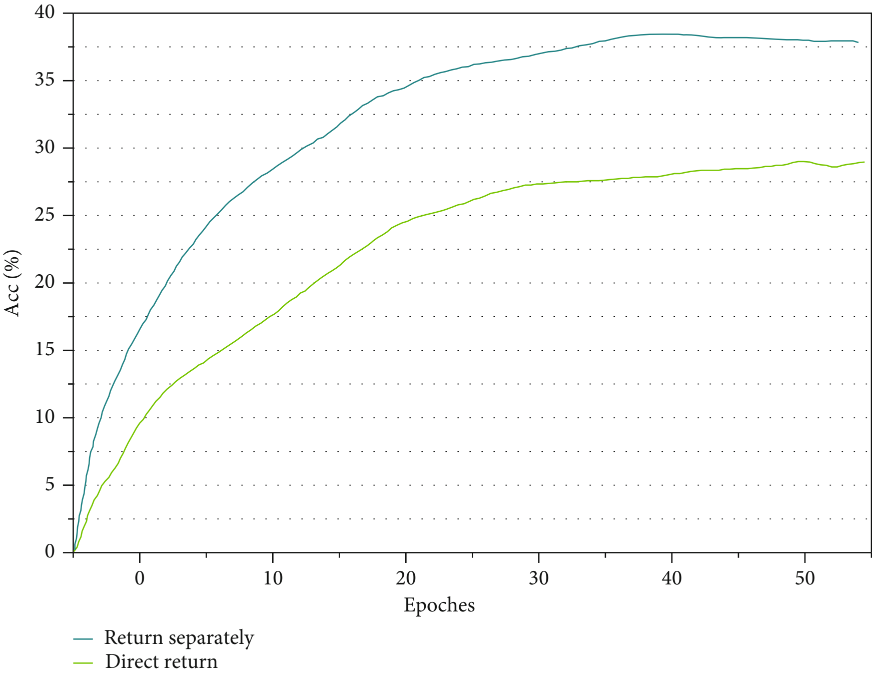

Figure 5: The trend of offset vector regression accuracy with training rounds on the validation set.

To reduce the difficulty of offset vector regression, this paper decomposes the offset vector regression task into two subtasks, unit vector regression, and vector modal length regression. To verify the effectiveness of this strategy, direct and indirect regressions are performed on the offset vectors, respectively. Figure 5 shows the variation curve of Acc-1 with the number of training rounds (epoch) of the offset vector regression accuracy on the validation set of the hard-pose subset of the UBC3V dataset. The strategy of decomposing the offset vector regression task significantly improves the accuracy of offset vector regression. The reason for the unsatisfactory results of direct regression of offset vectors mentioned in the literature is that the lengths of offset vectors from surface points to joint points vary greatly from one part to another, which makes the regression target have a large variance, and offset vectors with larger lengths dominate the training loss, resulting in a more difficult training. The unit offset vectors have the same modal length; so, the measures taken in this paper to regress the bit vectors and vector modal lengths separately avoid the defects brought by the direct regression of offset vectors and reduce the difficulty of offset vector regression.

We can shorten the time of the whole matching process by reducing the matching range at each matching, thus improving the efficiency of the whole matching algorithm while still taking into account the accuracy rate. At the same time, as the human body is in motion, there is a faster movement, when the direct use of the above algorithm will lead to a larger recognition error. All things considered, to ensure the reliability of the algorithm operation, every L frame, we research the whole sample space; that is, we can get a more accurate 3D pose.

To verify the reliability and feasibility of the whole interaction system, we conduct experimental verification by two indexes: accuracy and real-time. By comparing the response time of the proposed interaction method with the traditional point-of-view gaze interaction method, the reliability of the whole interaction system is illustrated and because the action recognition the new VR interaction method has its advantages because the interaction method discards the problems of fixed buttons and rigid interaction in the traditional interaction method. Since four common body movements and their corresponding command operations were proposed in the experimental design, the four body movements were investigated separately, as shown in Figure 6. Therefore, 50 interaction experiments were conducted for each of the four body movements, and the number of successful interactions and the average interaction time for each action were recorded. We can see that the accuracy rate of all the movements exceeded 70%, which shows the feasibility of VR interaction with body movements, but we can see that the accuracy rate of different movements varies greatly, and the recognition accuracy is better for the movements with larger amplitude and left-right expansion, while the recognition accuracy is worse for the movements with a smaller amplitude and front-back expansion, and the speed and frequency of the switching between different movements are lower. When switching between different actions, the speed and frequency are not easy to be too fast,;otherwise, it will easily lead to possible misoperation or reduce the accuracy of the command operation.

The 3D diffusion model human identification method has a better effect on the detection of difficult joints such as wrists, knees, and ankles. It is analyzed that this is because the difficult joints such as wrist, knee, and ankle are prone to occlusion and distortion, while the 3D diffusion model human recognition approach uses the similarity between global regions to assist in inference; so, it can improve the

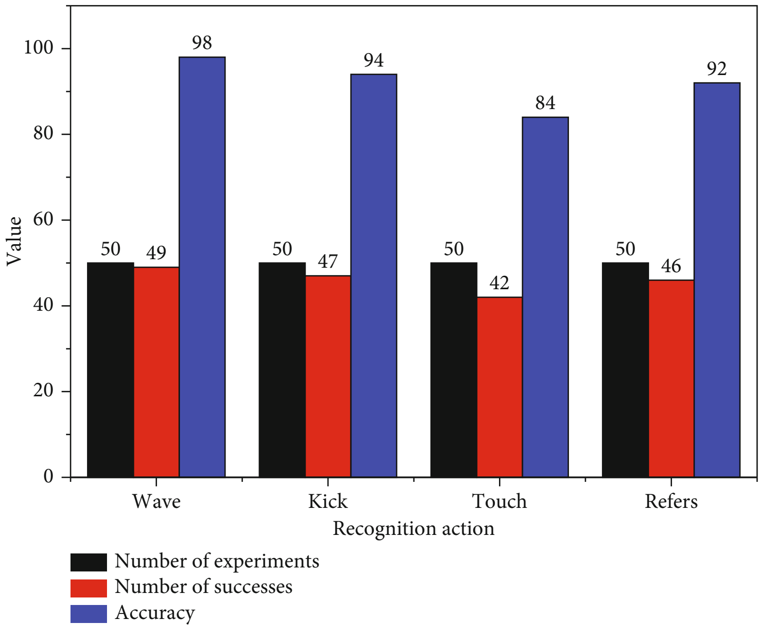

Figure 6: Experimental results of motion recognition accuracy.

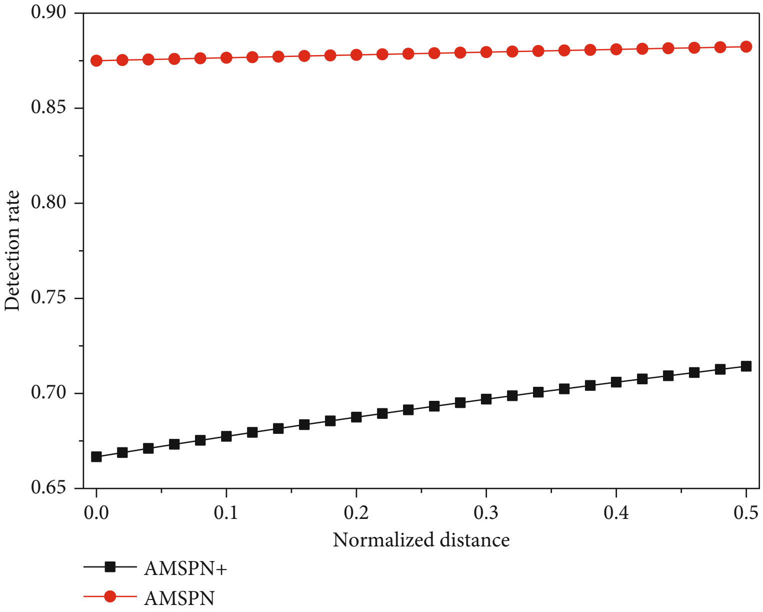

Figure 7: Average PCKh curves of the 3D diffusion model and the conventional model on the MPII dataset.

accuracy of these joints very well. Meanwhile, the average PCKh curves of the conventional model network without the 3D diffusion model human recognition approach (and the multi-scale pyramidal network based on the attention mechanism using the 3D diffusion model human recognition approach) are plotted, as shown in Figure 7.

Since the interaction system needs to pay attention to its real-time nature, otherwise it has a great impact on the interaction experience and reduces the interaction efficiency, we compared the average time of the VR interaction process based on the 3D diffusion model human recognition action with the traditional point-of-view gaze interaction process time, and the experimental results are shown in Figure 8.

The time is the average time after 50 experimental measurements, and the experimental results show that the recognition time of the human skeleton recognition method based on the 3D diffusion model is mostly shorter than that of the recognition method using the point-of-view gaze system, which illustrates the advantages of the 3D diffusion model recognition method of the human skeleton proposed in this paper. Moreover, because the point-of-view gaze interaction system needs to arrange fixed operation buttons in VR

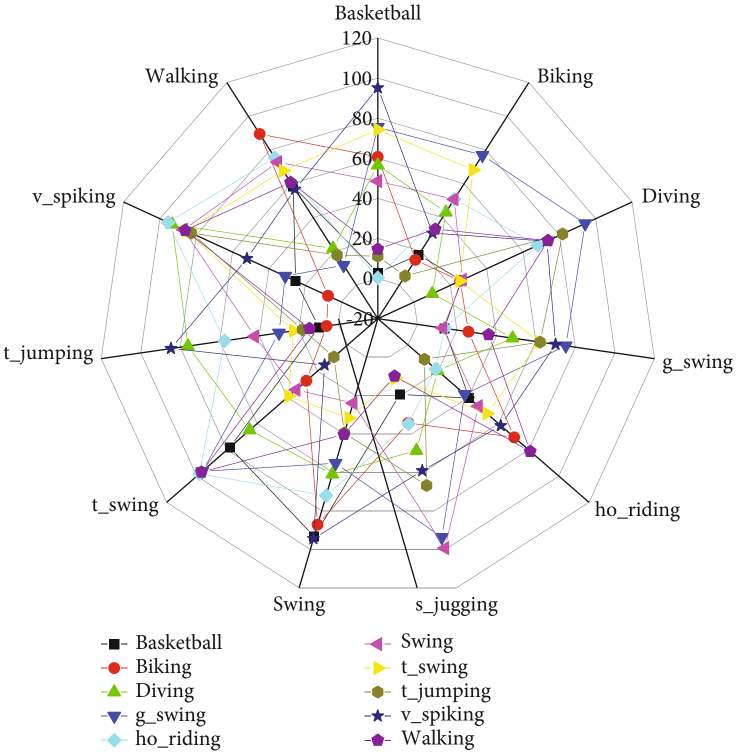

Figure 8: Action recognition interaction time comparison.

space, it is extremely inconvenient to use and affect the immersive experience of the whole VR space, while the interaction method based on 3D diffusion model action recognition perfectly avoids these shortcomings and makes the whole interaction process smoother and more efficient.

5. Conclusion

Thanks to the unique advantages of the 3D diffusion model itself, the development of point cloud acquisition technology, and the enhancement of hardware computing power, the research on human skeleton detection and extraction has become a new research hotspot in recent years. As a core task in the field of computer vision and computer graphics, the detection description and model alignment of key points have made some research progress, but there are still a lot of problems to be solved. In this paper, we propose a complete set of keypoint detection, key point description, saliency region detection, and model alignment algorithms for human skeleton detection and extraction from local features. This paper mainly does the following work: (1) proposes to build a skeleton point detection model using an improved bottom-up scheme, which first detects all the skeleton point positions of the human body in the picture, and then reorganizes individual instances according to the association information, and the whole picture only needs to be entered once from the prediction network, thus eliminating the impact of uncertainty of the human body; (2) proposes to use the existing skeleton point and (2) propose to use the existing skeleton points to construct the association information between the skeleton points as a new feature to be provided to the CNN for training, so that the association information between the skeleton points can be obtained as a detection problem; and (3) use the multiscale equalization module to equalize the features of different scales separately and dynamically assign different attention weights to the features of different scales according to the loss function when detecting different joints, so that the features of different scales can be used more. The features at different scales are dynamically assigned different attention weights according to the loss function when detecting different nodes so that the features at different scales can be used more efficiently.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the University-Level Foundation Project of Guilin University of Aerospace Technology: Research on Design of the Way of Expression in Virtual Space from the Perspective of Semiotics, No. XJ20KT22.

References

[1] L. Chen, N. Ma, P. Wang et al., “Survey of pedestrian action recognition techniques for autonomous driving,” *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 458–470, 2020.

[2] Z. Liu, Z. Lin, X. Wei, and S. C. Chan, “A new model-based method for multi-view human body tracking and its application to view transfer in image-based rendering,” *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1321–1334, 2018.

[3] H. Rahmani, A. Mian, and M. Shah, “Learning a deep model for human action recognition from novel viewpoints,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 667–681, 2018.

[4] A. Nadeem, A. Jalal, and K. Kim, “Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy markov model,” *Multimedia Tools and Applications*, vol. 80, no. 14, pp. 21465–21498, 2021.

[5] P. Pareek and A. Thakkar, “A survey on video-based human action recognition: recent updates, datasets, challenges, and applications,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021.

[6] W. Yang, L. Qingtang, H. Haoyi et al., “Personal active choreographer: improving the performance of the Tujia hand-waving dance,” *IEEE Consumer Electronics Magazine*, vol. 7, no. 4, pp. 15–25, 2018.

[7] M. Ullah, M. Mudassar Yamin, A. Mohammed, S. Daud Khan, H. Ullah, and F. Alaya Cheikh, “Attention-based LSTM network for action recognition in sports,” *Electronic Imaging*, vol. 2021, no. 6, pp. 302-1–302-6, 2021.

[8] J. S. Im and J. H. Kim, “A quantification method of human body motion similarity using dynamic time warping for key-points extracted from video streams,” *Journal of IKEEE*, vol. 24, no. 4, pp. 1109–1116, 2020.

[9] S. K. Yadav, A. Singh, A. Gupta, and J. L. Raheja, “Real-time yoga recognition using deep learning,” *Neural Computing and Applications*, vol. 31, no. 12, pp. 9349–9361, 2019.

[10] N. T. Thành, L. V. Hùng, and P. T. Công, “An evaluation of pose estimation in video of traditional martial arts presentation,” *Journal of Research and Development on Information and Communication Technology*, vol. 2019, no. 2, pp. 114–126, 2019.

[11] O. AlShorman, B. Alshorman, and M. S. Masadeh, “A review of physical human activity recognition chain using sensors,” *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 8, no. 3, pp. 560–573, 2020.

[12] H. T. Chen, Y. Z. He, and C. C. Hsu, “Computer-assisted yoga training system,” *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 23969–23991, 2018.

[13] T. Singh and D. K. Vishwakarma, “Video benchmarks of human action datasets: a review,” *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1107–1154, 2019.

[14] H. Hegazy, A. Nabil, M. Abdelsalam et al., “Usability study of a comprehensive table tennis AR-based training system with the focus on players' strokes,” *Journal of Ubiquitous Systems & Pervasive Networks*, vol. 13, no. 1, pp. 1–9, 2020.

[15] A. Sharif, M. A. Khan, K. Javed et al., “Intelligent human action recognition: a framework of optimal features selection based on Euclidean distance and strong correlation,” *Journal of Control Engineering and Applied Informatics*, vol. 21, no. 3, pp. 3–11, 2019.

[16] X. D. Li, Y. L. Wang, Y. He, and G. Q. Zhu, “Research on the algorithm of human single joint point repair based on Kinect,” *Techniques of Automation & Applications*, vol. 35, no. 4, pp. 96–98, 2016.

[17] P. Girdhar, “Vision based human activity recognition: a comprehensive review of methods & techniques,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 10, pp. 7383–7394, 2021.

[18] M. Sakamoto, T. Shinoda, K. Sakoma, T. Ishizu, A. Takei, and T. Ito, “Interactive projection mapping using human detection by machine learning,” *Journal of Advances in Artificial Life Robotics*, vol. 1, no. 1, pp. 85–89, 2020.

[19] N. Eichler, H. Hel-Or, I. Shimshoni, D. Itah, B. Gross, and S. Raz, “3D motion capture system for assessing patient motion during Fugl-Meyer stroke rehabilitation testing,” *IET Computer Vision*, vol. 12, no. 7, pp. 963–975, 2018.

[20] T. T. Nguyen, V. H. le, D. L. Duong, T. C. Pham, and D. le, “3D human pose estimation in Vietnamese traditional martial art videos,” *Journal of Advanced Engineering and Computation*, vol. 3, no. 3, pp. 471–491, 2019.

[21] H. Wang and L. Wang, “Beyond joints: learning representations from primitive geometries for skeleton-based action recognition and detection,” *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4382–4394, 2018.