applied sciences

MDPI

Article

Motion Capture for Sporting Events Based on Graph Convolutional Neural Networks and Single Target Pose Estimation Algorithms

Chengpeng Duan [1,*], Bingliang Hu [1], Wei Liu [2] and Jie Song [2]

[1] Xi'an Institute of Optics and Precision Mechanics of Cas, University of Chinese Academy of Sciences, Xi'an 710119, China
[2] Xi'an Institute of Lead Ir Technology Co., Ltd., Xi'an 710119, China
* Correspondence: duanchengpeng@opt.ac.cn

Abstract: Human pose estimation refers to accurately estimating the position of the human body from a single RGB image and detecting the location of the body. It serves as the basis for several computer vision tasks, such as human tracking, 3D reconstruction, and autonomous driving. Improving the accuracy of pose estimation has significant implications for the advancement of computer vision. This paper addresses the limitations of single-branch networks in pose estimation. It presents a top-down single-target pose estimation approach based on multi-branch self-calibrating networks combined with graph convolutional neural networks. The study focuses on two aspects: human body detection and human body pose estimation. The human body detection is for athletes appearing in sports competitions, followed by human body pose estimation, which is divided into two methods: coordinate regression-based and heatmap test-based. To improve the accuracy of the heatmap test, the high-resolution feature map output from HRNet is used for deconvolution to improve the accuracy of single-target pose estimation recognition.

Keywords: single-target pose estimation; graph convolutional neural network; deep learning; target detection

Citation: Duan, C.; Hu, B.; Liu, W.; Song, J. Motion Capture for Sporting Events Based on Graph Convolutional Neural Networks and Single Target Pose Estimation Algorithms. *Appl. Sci.* **2023**, *13*, 7611. https://doi.org/10.3390/app13137611

Academic Editor: Antonio Fernández-Caballero

Received: 20 April 2023
Revised: 16 June 2023
Accepted: 16 June 2023
Published: 27 June 2023

1. Introduction

Computer vision (CV) is a thriving research area in machine learning, fueled by the exponential growth of visual information and image data in various modern applications. The advancements in deep neural network models and image processing techniques have led to the widening of the applicable domain scenarios and the definition of new problems in CV [1]. Human pose estimation, as a significant area in computer vision, has wide-ranging applications in various fields. For instance, in the field of human-computer interaction, customized gestures or movements can enable humans to control computers [2]; in the entertainment gaming industry, players can experience immersive gameplay by physically controlling game characters [3]; in sports analysis, pose estimation systems can provide insights into the technical movements of athletes from multiple perspectives [4]; in autonomous driving, posture estimation can analyze driver and passenger behavior and provide early warnings of abnormal driving conditions [5].

Convolutional neural networks, commonly used for image processing in deep learning, can also be applied to graph data processing. Kpif et al. [6] proposed graph convolutional networks (GCNs), which introduced a novel approach to action recognition through skeleton information. Yoon et al. [7] were the pioneers in applying graph convolutional networks to action recognition and developed the ST-GCN method for behavior classification, demonstrating outstanding results on two large action data sets. Additionally, Simonovsky et al. [8] designed a graph edge convolutional neural network (GECNN) that fused the results of the graph edge convolutional neural network and graph convolutional neural

network, yielding significant improvements in the performance of graph convolutional networks, computer vision, and human pose estimation are rapidly advancing fields that have numerous applications in various domains. With the emergence of deep neural network models and image processing techniques, the applicable domain scenarios are broadening, and new research problems are being defined. The application of convolutional neural networks in graph data processing, such as graph convolutional networks and graph edge convolutional neural networks, has provided novel solutions to the challenges in action recognition and pose estimation.

Human pose estimation is the process of detecting and locating the coordinate position and orientation of human joints from images or videos. Human pose estimation plays a very important role in areas such as action recognition [9], intelligent recognition [10], and human-computer interaction [11]. It is in sports, as an important application area for the application of human posture estimation, that can help to effectively assess the overall ability of athletes in multiple dimensions by analyzing their movements and providing an informed reference for athletes; research on body pose estimation in sports is developing very rapidly, with algorithms evolving from the process of analysis and modeling of sporting events.Research on body pose estimation in sports is developing very rapidly, with algorithms evolving from the original traditional methods of research to the current deep learning approach [12]. Using neural networks, deep learning algorithms for human body pose estimation models can extract sufficient target image features for accurate identification and detection of Sports video analysis-based training and match assistance systems that have a wide range of applications both at home and abroad [13]. The Coach's Eye application, developed by TechSmith, records the state of the athlete through a mobile device or camera and plays the video in slow motion. The Coach's Eye application, developed by TechSmith, records the state of the athlete through a mobile device or camera, plays the video in slow motion and analyses it frame by frame, analyses comparative movements side-by-side, uses advanced analysis tools such as stopwatches to mark time, The system, due to the use of slow motion to analyze movements, is analysis efficiency needs to be improved [14]; STATS' SportVU (Player Tracking Analysis System) in the USA was one of the first systems to apply video analysis technology to sports competitions [15]. In the domestic sports video analysis system, Chuangbing DA-TA slices the second-level data of each game, and uses the distributed computing platform to perform multi-dimensional statistics on the cloud data, which is more accurate in the game data statistics, but lacks the player's action analysis [16]; the data acquisition and analysis system of Lingxin sports events uses the football and player trajectory data provided by the high-speed camera to collect the player's movement information, and uses the template-matching method based on the Lingxin sports system to track the players. The system records the data related to the player's movement. Through the data analysis of the player's performance, in the football game, the accuracy of the individual player's action analysis needs to be improved [17].

Pose estimation can be divided into single-pose estimation and multi-person pose estimation. Single-pose estimation predicts the body parts and joints of the human body in the image that has been cropped and processed. In contrast, multi-person pose estimation can be divided into top-down searches. In the early methods of single-body pose estimation, the body parts were detected by using the artificially extracted Histogram. In the early methods of single-body pose estimation, the body parts were detected by using the artificially extracted Histogram of Oriented Gradient (HOG) features, and the human pose structure was represented by using the probability model map (rather than the tree model) [18]. After Toshev et al. [19] proposed DeepPose using AlexNet to regress the coordinates of human node locations to proceed, deep neural network-based models rapidly occupied the field of pose estimation. Tompson et al. [20] learned human pose structure by combining depth features and graphical models. Carreira et al. [21] proposed an iterative error feedback method (Iterative Error Feedback, IEF) to train Convolution Neural Networks (CNNs), which will repeatedly feed the input into the network and self-correct the network based on the prediction feedback. In recent years, the application

of single-body pose estimation to athletes is mainly in the analysis of athletes' movements during competition. Farrukh et al. [22] proposed a yoga training system by integrating computer vision techniques, and the system analyzes the trainer's pose by extracting body contours, skeleton, dominant axes, and feature points from the front and side views [23]. Then, based on the domain knowledge of yoga training, visualized posture correction instructions were proposed with high accuracy in correcting the trainer's posture. Asha et al. [24] used a deep key frame extraction method for analyzing weightlifting sports training videos in order to monitor and analyze athletic poses of athletes training in professional sports, and the proposed DKFE outperformed the comparison methods in terms of key pose probability estimation and key pose extraction [25].

Pose estimation has become a popular research topic in sports movement analysis. In this stage, researchers are using single static images as input to integrate similar images in neighboring pixel points and complete the local search of human parts through superpixels [26]. They apply the variability part model to achieve human part recognition, which reduces the interference of background on part recognition. Additionally, the supervised training convolutional network provided by the ordinal depth based on single human joints can effectively improve the accuracy of remote mobilization detection in the scene.

This paper specifically addresses the following problems based mainly on graph neural networks and single target pose estimation algorithms. The contributions of this paper are as follows:

(1) The model combining graph neural network and HRNet can effectively improve the accuracy and efficiency of pose estimation tests. This can help accurately identify the limb movements of athletes in the game and provide an effective reference for athletes' movement analysis.
(2) Based on the end-to-end deep network for personnel detection and deconvolution operation through multiple upsampling and HR-Net, the shallow layer in the fusion neural network for human location information recognition improves single target recognition. This can circumvent some inspection errors caused by overlapping or shading and, to a certain extent, reduce the occurrence of missed and wrong inspections.
(3) The graph neural network-based human pose estimation method is further optimized on the traditional graph structure-based method. This ensures adequate extraction of human pose features while reducing the time required for extraction and further improving the accuracy and efficiency of object detection and human pose estimation.

This paper aims to investigate the action recognition of athletes in sports competitions using the skeletal point dataset and graph neural network. The paper begins by introducing the current development of computer vision and highlighting the specific application of single-body pose estimation in sports. The purpose and significance of the study are also explained. In the second part of the paper, related work in the field is reviewed, and some of the most commonly used single-body pose estimation algorithms are analyzed. The third part of the paper introduces the algorithms used in this study and presents a detailed technical roadmap for the entire paper. The fourth part of the paper describes the experimental process, which involves applying the pose estimation algorithm to athlete photo sampling recognition on the coco dataset, and verifying the validity and applicability of the model on the validation set. Additionally, the performance of the proposed model is compared with some of the current mainstream algorithms, and a comprehensive prediction result is presented. Towards the end of the paper, some advantages and disadvantages of the proposed model are discussed, and the entire study is summarized. Finally, future work is outlined.

Overall, this paper follows a standard structure for a research study in computer vision, which includes an introduction, related work, methodology, experiments, results, and conclusions. The language used is academic, and the presentation of the content is clear and concise.

2. Related Work

Human pose estimation, which can also be called human keypoint detection, is a relatively basic task compared to other fields in machine vision, such as target detection, image segmentation, image enhancement, image generation, and face classification recognition, whose main purpose is to estimate the human pose by correctly connecting the detected human joints in the picture [27,28]. According to whether it contains three-dimensional depth information, human keypoint detection algorithms can be divided into 2D keypoint detection and 3D keypoint detection, and in 2D and 3D keypoint detection can be divided into single and multi-person keypoint detection. In 2016, Wei et al. merged convolutional networks into a pose machine to learn image features and image-dependent spatial models for pose estimation [29], and Newell et al. proposed hourglass-type convolutional networks [30]. Both methods perform the process of detecting people first and then key points for each person. In 2017, a researcher proposed the OpenPose measure, which first detects all keypoints in an image and then associates keypoints in groups by affinity to obtain the corresponding individual person. In 2019, a researcher proposed an improved version of OpenPose, proposing a real-time method for detecting 2D poses of multiple people in an image, which learns how to associate body parts with individuals in the image. OpenPose is applicable to both single- and multi-person scenes and has both better robustness and real-time performance. Compared with 2D skeleton points, 3D skeleton points contain richer skeleton features, such as stereo angle information between joints and depth information around bones.

There are three main methods to obtain 3D skeletal points. The first one is to obtain the (x, y, z) coordinates of the joints by the subject wearing a motion tracking sensor. This method is accurate and robust in estimating results, and does not introduce estimation errors due to background occlusion. However, the data acquisition process is tedious, and the sensor equipment is expensive. The second one uses a depth camera and a color camera with a skeletal point estimation algorithm for estimation [31], which is more accurate and easier to use. However, it is sensitive to occlusion and requires the use of algorithms to process the occluded points that are incorrectly predicted, and the portability and higher price affect the large-scale use of depth cameras. The third one is RGB video-based pose estimation, which only needs a general camera to obtain 3D skeletal points, the method is more suitable for large-scale use and promotion, but the accuracy rate is currently inferior to the second method because of the lack of accurate depth information. The RGB video-based human pose estimation algorithm is a hot topic at present and has the potential for large-scale application, so the third method is used in this thesis to extract 3D skeletal points. Based on RGB video 3D key point extraction, the early method of convolutional neural network was used to estimate 3D pose directly from RGB images, and then with the development of deep learning technology, many methods for 3D pose estimation were generated. 2019 Pavllo et al. proposed a full convolutional model for 3D human pose prediction [32], which only needs to do temporal convolution of 2D key points to achieve human 3D keypoint prediction. Erik et al. proposed PoseDRL [33], a trainable active pose estimation architecture based on deep reinforcement learning, which has high performance for pose estimation in multi-person complex scenes. The analyzed and compared VideoPose3D model can achieve high accuracy with fewer parameters and can be trained and inferred quickly.

Human keypoints are the most important information when describing human pose and detecting human movement. As shown in Figure 1 below, Keypoints Detection is the main technique for the human pose estimation task, which aims to accurately find the location of each key node of the human body contained in an image or a video frame. The number of keypoints is determined by the dataset used, and the number of keypoints and their locations vary slightly from dataset to dataset.

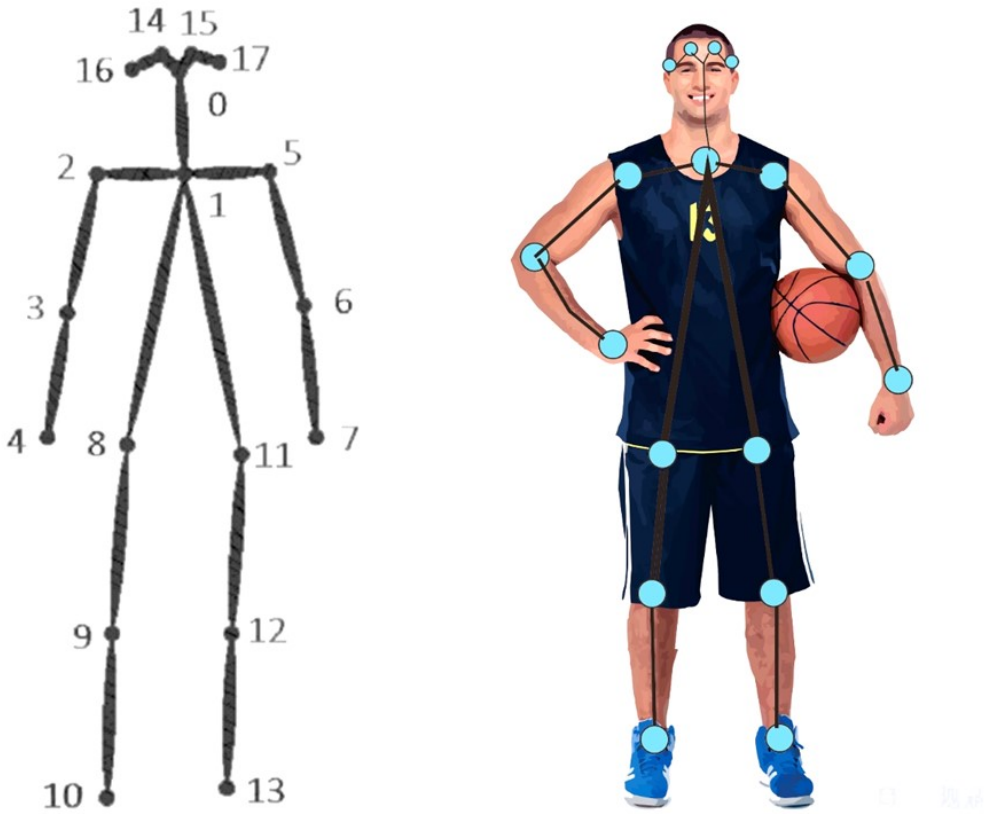


Figure 1. Human skeleton diagram composed of key points.

Recently, graph convolutional neural networks (GCNNs) have become increasingly popular due to their ability to model local structures and node dependencies on graphs. The first GCNN was proposed by vila et al. in 2013, which defined graph convolution in the spectral space based on graph theory [34]. However, the initial spectral method had the drawback of high time and space complexity [35]. Subsequent methods such as ChebNet and GCN parametrized the convolution kernel in the spectral method [36], which greatly reduced the space-time complexity.

Inspired by these methods, spatial methods were applied, and attention mechanisms and serialization models were used to model the weights between nodes in the node domain. With the improvement in convolutional operators, researchers began to consider various graph characteristics such as how to model higher-order information on the graph, and fine-grained design for graphs with features on edges and heterogeneous graphs. In addition, the scalability of models to large-scale graphs and the speed of training have also received wide attention. In 2019, Sun et al. proposed a High-Resolution Network (HRNet), a network model that distinguishes itself from previous low-resolution networks by connecting convolutional group networks of different resolutions together in a parallel way. This network uses parallel connections to maintain the input resolution from beginning to end and allows for multi-scale information fusion. By fusing feature information from different resolutions, HRNet effectively extracts feature information from the target and allows for better estimation of a single pose. In human pose estimation, the general process of HRNet involves target detection, which is used to detect the approximate location of the human body [37]. This location is then used to detect all keypoints of the human body, and a regressor is used to predict the individual key locations. After the estimation of all keypoints, the key points are returned to the original map in the form of a heatmap, achieved by adding a convolutional layer to the last layer of the high-resolution network model. This converts the key point locations into a full-resolution heatmap [38].

Overall, HRNet represents an advancement in GCNNs for human pose estimation, allowing for effective multi-scale information fusion and feature extraction. Future research may focus on further improving the scalability and efficiency of GCNNs for large-scale graph [39].

3. Method

This paper proposes an algorithmic framework for motion capture, which includes several parts as depicted in Figure 2. Firstly, a human pose estimation algorithm is utilized to extract the skeletal information from the video. Specifically, this paper adopts a graph neural network-based algorithm for human action recognition, which models the skeletal data into a fixed topological graph structure to extract spatiotemporal features of the skeleton using graph convolutional networks, including both graph convolutional neural networks and adaptive graph convolutional networks [40].

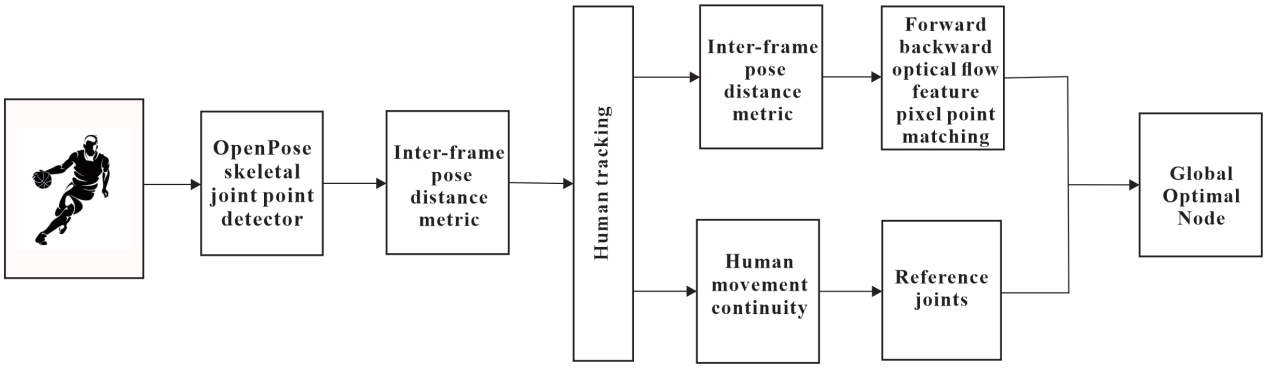


Figure 2. Framework diagram of human posture estimation.

Secondly, a video-based human pose estimation algorithm is proposed. This algorithm consists of several steps. Firstly, the human pose is estimated using the OpenPose algorithm for still images. Then, a human tracking model is established based on inter-frame pose distance metric, and UltraPixel segmentation is performed on the images to locate the UltraPixel where skeletal joints are located. The UltraPixel is connected with the box centered on the joint point, and the intersection between the UltraPixel and the box is considered as the minimum candidate joint point set. Forward and backward searching for better joint points is conducted based on optical flow and human motion continuity to establish the candidate joint point set and the reference joint point [41]. Finally, the optimal global human pose is generated for each frame by reorganizing these body parts, and the key points with lower confidence are generated in each frame to generate a better candidate joint point set [42].

3.1. OpenPose-Based Human Pose Estimation Algorithm

The OpenPose algorithm represents the latest advances in human pose estimation for static images. This two-dimensional pose estimation algorithm is highly effective at detecting multi-target human bodies in images [43]. One of its key advantages is its bottom-up approach, which allows it to efficiently identify and track individuals even in images with a large number of people [4]. At the heart of this algorithm is an explicit non-parametric representation of skeletal joint point associations that encodes the orientation and position of human limbs [44]. This representation is used to detect jointly learned body parts and to identify associations between them. The algorithm also employs a set of frameworks for detecting body part associations and employs greedy analysis algorithms to estimate human body pose accurately [45]. Taken together, these features enable OpenPose to achieve outstanding performance in human pose estimation [46]. The algorithm's ability to detect multiple individuals in a single image, its high accuracy, and its versatility make it a valuable tool for a wide range of applications, including computer vision, robotics, and human-computer interaction.

Figure 3 illustrates the complete network architecture of OpenPose, which comprises two distinct branches. The upper branch network, depicted in orange, is responsible for producing the confidence map L, which indicates the likelihood of the presence of skeletal joints. The lower branch network, shown in blue, is responsible for generating the partial affinity vector field S [47]. Prior to processing, the input image undergoes convolutional neural network (CNN) processing, which involves the use of the first 10 layers of VGG-19 for initialization and fine-tuning. This step allows for the extraction of a collection of feature maps F.

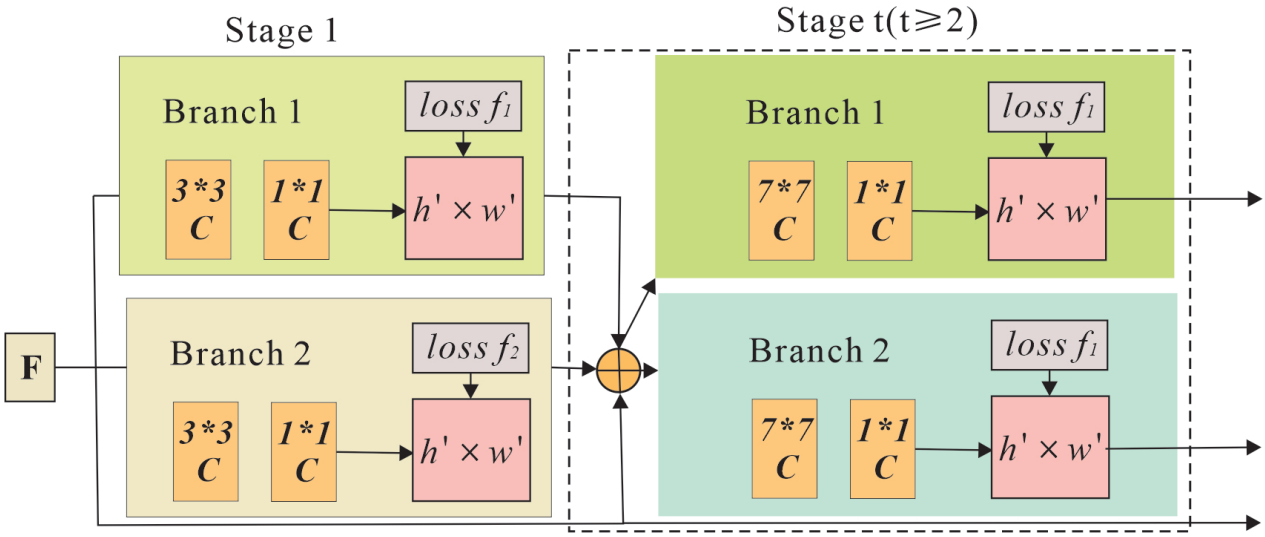


Figure 3. OpenPose network architecture diagram.

As shown in Figure 3, the general flowchart of the OpenPose algorithm, the algorithm requires an input color image of size $w \times h$ (Figure 4a), and first predicts the set of part confidence maps S (as in Figure 4b) from the input image through a feedforward network, and also predicts the part affinity vector fields L (Part Affinity Fields (PAFs) (Figure 4c), which encode the association between each skeletal joint in the human body, are predicted for subsequent association matching of skeletal joints. The set $(S_1, S_2, \ldots, S_J)$ denotes each skeletal joint point with J body part confidence maps, where $S_J \in R \times h$, $j \in 1 \ldots J$. Set the combined $L = (L_1, L_2, \ldots, L_C)$ with C site affinity vector fields for each limb site, where $L \in R \times h \times 2$, $c \in 1 \ldots C$, and L encodes a 2D vector for representing the position in the image [20]. Finally, the part confidence map and the part affinity vector field are resolved by the greedy algorithm to output 2D human skeletal joint points for all people in the image. Algorithm 1 is the pseudo-code of IEF, a generic framework for pose estimation used for feature extraction.

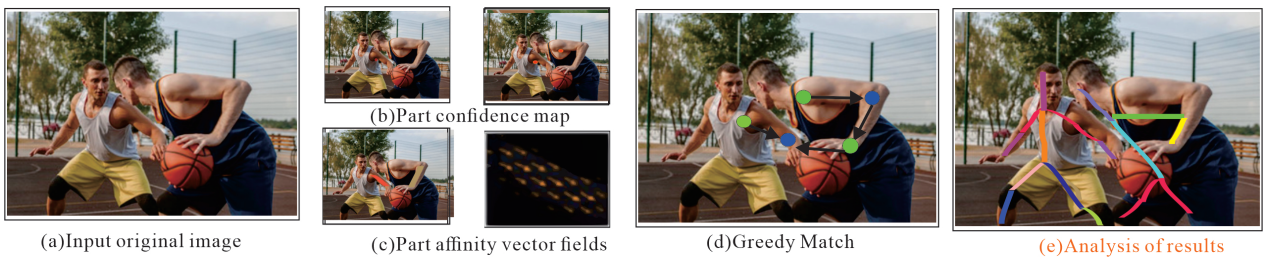


Figure 4. Flowchart of OpenPose algorithm.

Algorithm 1: Learning Iterative Error Feedback with Fixed Path Consolidation

```
procedure FPC-LEARN
    Initialize y_0
    E ← {}
    for t ← 1 to (T_steps) do
        for all training examples (I, y) do
            ε_t ← e(y, y_t)
            end for
        E ← E ∪ ε_t
    for j ← 1 to N do
        Update Θ_f and Θ_g with SGD, using loss and target corrections E
        end for
    end for
    end procedure
```

In the first stage, the network outputs a set of confidence maps $S = \rho l(F)$, and a set of partial affinity vector fields $L, l = \varphi l(F)$, where $\rho(F)$ and $\varphi l(F)$ are product of the Convolutional Neural Networks (CNN) inferred in the first stage, and each branch can be iterated. When iterating, the output of the branch network in the previous stage will be mapped with the features of the original image F as the input of the branch network in the current stage, which is used to produce more refined predictions. As in Equations, ρ and φ^t are the intermediate results of step t.

$$S^t = \rho^t\left(F, S^{t-1}, L^{t-1}\right), \forall t \geq 2 \tag{1}$$

$$L^t = \varphi^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2 \tag{2}$$

To enable iterative network structure and outcome prediction for each branch, a loss value is assigned at the end of each stage [48]. The upper and lower branch networks incorporate a loss function, which measures the discrepancy between the predicted and true values [49]. Spatial weighting of the loss functions is applied to address practical challenges, such as incomplete labeling in some datasets. Specifically, in the T-phase, the two branch networks utilize distinct loss functions. Refining the language and adopting a more formal tone yields the following revised paragraph:

To facilitate iterative network structure and outcome prediction for each branch, a loss value is assigned at the end of each stage. The upper and lower branch networks incorporate distinct loss functions, which quantify the dissimilarity between the predicted and ground truth values.

$$f_s^t = \sum_j^J \sum_P W(P) * ||S_j^t(p) - S_j^*(p)||_2^2 \tag{3}$$

$$f_L^t = \sum_{c=1}^C \sum_P W(P) * ||L_c^t(p) - L_c^*(p)||_2^2 \tag{4}$$

The OpenPose network consists of a body part confidence map, denoted as S^*, and a vector field of part affinities, denoted as L^*, which are associated with the true label. A binary mask W is applied to avoid error penalties in certain circumstances, such as when the network predicts part confidence maps and affinity vector fields for an image location p that lacks a label. This mask W is used to prevent penalties in such cases. To ensure that gradients do not vanish during the training process, the OpenPose network is supervised at each intermediate stage by periodically replenishing the gradients. The final objective

function is a combination of the body part confidence map and the vector field of part affinities, which is:

$$f = \sum_{t=1}^{T} (f_S^t + f_L^t) \tag{5}$$

3.2. Graph Convolutional Neural Network

Convolutional neural networks (CNNs) have achieved remarkable success in extracting local features from Euclidean data possessing regular spatial structures, such as in image analysis. The utilization of translation invariance, local connectivity, and image structure enables CNNs to extract shared local features from entire datasets. However, when applied to complex domain data without regular spatial structures, such as social networks in social sciences, distributed interconnected sensing networks, knowledge graphs, and networks of interacting protein molecules, CNNs are not as effective. This has led to an increasing focus on graph-structured convolutional neural networks (GCNNs) as a deep learning method suited for graph-structured data.

GCNNs involve the construction of convolutional and pooling operators. Convolutional operators aim to capture local node structures, while pooling operators aim to learn hierarchical representations and reduce parameters. For node-level tasks, the primary focus is on enhancing the expression of each node. In practical applications, pooling operators may not be required, and early work has concentrated on constructing convolution operators on the graph. Pooling operators are typically employed for graph-level tasks. The key challenge in GCNNs is developing convolution operators on the graph that capture the local connectivity of nodes. The graph structure determines the adjacency matrix, which represents node connectivity. Various convolution operators have been proposed, including spectral-based and spatial-based methods. In summary, GCNNs provide a powerful tool for analyzing complex domain data represented by graphs. The construction of convolution and pooling operators is critical to develop effective GCNN models. GCNNs have a broad range of applications, including social network analysis, protein interaction analysis, and knowledge graph analysis. A common CNN algorithm pseudocode example is presented in the following Algorithm 2.

Algorithm 2: ID-GNN embedding computation algorithm

Input: Graph $\mathcal{G}(\mathcal{V};\mathcal{E})$, input node features $\{x_v, \forall v \in \mathcal{V}\}$; Number of layers K trainable functions $\mathrm{MsG}_1^{(k)}(\cdot)$ for nodes with identity coloring, $\mathrm{MSG}_0^{(k)}(\cdot)$ for the rest of nodes; $\mathrm{EGO}\,(v,k)$ extracts the K-hop ego network centered at node v, indicator function $\mathbb{1}[s=v]=1$ if $s=v$ else 0

Output: Node embeddings $\mathbf{h}_v$ for all $v \in \mathcal{V}$

$v \in \mathcal{V}$ do
  $\mathcal{G}_v^{(K)} \leftarrow \mathrm{EGO}\,(v,K)$, $\mathbf{h}_u^{(0)} \leftarrow \mathbf{x}_u, \forall u \in \mathcal{G}_v^{(K)}$
  for $k = 1, \ldots, K$ do
    for $u \in \mathcal{G}_v^{(K)}$ **do**
  $\mathbf{h}_u^{(k)} \leftarrow \mathrm{AGG}^{(k)}$
  $\mathbf{h}_v \leftarrow \mathbf{h}_v^{(K)}$

Existing graph convolutional neural networks are divided into two categories: spectral methods, which use the convolution theorem on the graph to define the graph convolution from the spectral domain, and spatial methods, which start from the node domain and aggregate each central node and its neighboring nodes by defining an aggregation function. The absence of translation invariance on the graph poses difficulties in defining convolutional neural networks in the nodal domain, and spectral methods use the convolution theorem to define the graph convolution from the spectral domain, and we first give the

background knowledge of the convolution theorem convolution theorem: the Fourier transform of the signal convolution is equivalent to the product of the Fourier transforms of the signal:

$$F(f * g) = F(f) * F(g) \tag{6}$$

where f, g denotes the two original signals, $F(f)$ denotes the Fourier transform off, - denotes the product operator, and * denotes the convolution operator. Doing the Fourier inverse transform on both ends of equation, we can obtain:

$$f * g = F^{-1}(F(f) * F(g)) \tag{7}$$

where $F^{-1}(f)$ denotes the Fourier inverse transform of the signal f. Using the convolution theorem, we can multiply the signal in the spectral space and then use the Fourier inverse transform to convert the signal to the original space to realize the graph convolution, thus avoiding the difficult problem of defining the convolution caused by the graph data not satisfying the translation invariance. The Fourier transform on the graph depends on the Laplace matrix on the graph, and in the following, we will give the definition of the Fourier transform on the graph. The definition of the Fourier transform on the graph depends on the eigenvectors of the Laplace matrix. Using the eigenvectors as a set of bases under the spectral space, the Fourier transform of the signal x on the graph is: $x = U\hat{x}$. Using the Fourier transform and the inverse transform on the graph, we can implement the graph convolution operator based on the convolution theorem:

$$x *_G y = U((U^T x)) \odot (U^T y) \tag{8}$$

where G^* denotes the graph convolution operator, x, y denotes the signal of the node domain on the graph, and $\odot$ refers to the Hadamard multiplication, which denotes the multiplication of the corresponding elements of two vectors. We replace the vector $U^T y$ by a diagonal array g_θ, then Hadamard multiplication can be transformed into matrix multiplication. By acting the convolution kernel $g\theta$ on the signal, the graph convolution can be expressed in the following form: Ug_θ, $U^T x$. The convolution theorem provides a way to define convolution on a graph by Fourier transform. Based on this definition of the convolution operator, a number of graph convolution neural networks have emerged at home and abroad.

4. Experiment

4.1. Introduction to the Dataset

In this paper, we chose the COCO dataset for training and validation. The COCO dataset is a dataset specifically used for object detection, human keypoint detection, semantic segmentation, and other research experiments. In this paper, 118,287 human pose images from COCO2017 were used as the training set for training, 5000 as the validation set for verification. There are 17 annotated keypoints for a human instance in the COCO dataset, and the annotation order is shown in Table 1 .

In addition, a number of images of standard basketball games were collected for calibration tests for stance identification and movement analysis of players in standard basketball games, as shown in Figure 5. The footage of a regular NBA game, with the following footage of the real part of the game, had 10 players on the court and each player in the frame was of medium size. This paper focused only on the stance of the players, so when marking the test set joints and bounding boxes, only the players were marked, not the spectators and referees, while masking marks were added to the spectators and referees in the training set, and no losses were counted in the masking range. The images currently available for calibration are around 2000, of which the test data is 400, and the image below shows an intercepted video of an NBA game.

Table 1. The dataset characteristics on COCO VAL 2017.

Marking Number	Name	Marking Number	Name
0	Nose	9	Right knee
1	Neck	10	Right ankle
2	Right shoulder	11	Left hip
3	Right elbow	12	Left knee
4	Right wrist	13	Left ankle
5	Left shoulder	14	Right eye
6	Left elbow	15	Left eye
7	Left wrist	16	Right ear
8	right hip	17	Left ear

Figure 5. Screenshot of the NBA game used for this research.

4.2. Experimental Platform

In this paper, we used Pytorch 1.10.0 deep learning framework, the operating system environment was Windows 10, the GPU was GTX 1080 Ti, the system memory size was 12 G, and the programming language was Python 3.8.1. The initial learning rate was 0.001, which decreases to 0.0001 at 200 rounds and decays to 0.00001 at 260 rounds. A total of 260 rounds of training were performed with a batchsize of 32 per GPU. Since the images in the dataset are not of the same size, image pre-processing is used to modify the images. In this paper, we experimented to crop the images in the COCO dataset to 256 × 192 size, and then achieved data enhancement by random flipping and random scaling.

4.3. Evaluation Criteria

The performance evaluation of the model in this paper is based on the Object Keypoint Similarity (OKS) evaluation metric. The OKS measures the similarity between the predicted

keypoint locations and the ground truth keypoint locations. The OKS is defined as follows:

$$OKS = \frac{\sum_i \exp(\frac{-d_i^2}{2s^2k_i^2})\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{9}$$

where d_i represents the Euclidean distance between the detected keypoint and the corresponding ground truth label, v_i is an indicator variable that denotes whether the keypoint is visible or not, s is the target scale size, and k_i is a constant controlling the decay of the OKS for each keypoint.

The paper reports several performance metrics, including the Average Precision (AP) and Average Recall (AR) scores. Specifically, AP^{50} denotes the AP score at OKS = 0.50, AP^{75} denotes the AP score at OKS = 0.75, and AP denotes the average score of AP at various OKS thresholds ranging from 0.50 to 0.95, as well as other 10 positions. Additionally, AP^M and AP^L denote the AP scores for medium and large-sized human targets, respectively. The AR score is calculated similarly to the AP score, but is based on recall instead of precision. The AR score is reported for OKS thresholds ranging from 0.50 to 0.95, as well as other 10 positions. In summary, the OKS evaluation metric and the various performance metrics reported in this paper provide a comprehensive and rigorous evaluation of the model's performance on human pose estimation.

4.4. Analysis of Experimental Results

For the recognition of a single person, this study used the multi-person pose method, whose target-tracking method is a variant of SiamRPN, and in practice the tracking point was more fixed relative to the target, but the size of the bounding box from which it obtained the target was not constant and sometimes. As shown in Figure 6 below, the tracking was not able to keep up due to the larger movements.

(a) Normal condition (b)Running condition (c)Passing condition (d)Throwing condition

Figure 6. Single-tracking detection frame that does not contain some key points.

Figure 7 illustrates the impact of bounding box size on pose-based motion recognition. Specifically, in the normal situation shown in Figure 7a, all movement joints were included in the bounding box. However, during running, passing, and throwing, as shown in Figure 7b–d, the tracker might fail to include all joints, thereby compromising the accuracy of pose-based motion recognition. To address this issue, this paper proposes a fixed bounding box size of 256 * 192 centered on TrackPoint, which reduces image size variation relative to network input size, enhances model input efficiency, and speeds up pose estimation. When there are multiple people, a multiple person approach is used, which avoids the significant impact of using a single person estimation method. However, this approach raises the problem of multiple people wrapping up, which requires identifying the target being tracked. To address this issue, the target is identified as the person in the centre of the box with the fullest number of joints and the largest relative response size, and the response values are accumulated for the heatmap position corresponding to each person's joints in the box. This approach enhances the accuracy and reliability of pose-based motion recognition.

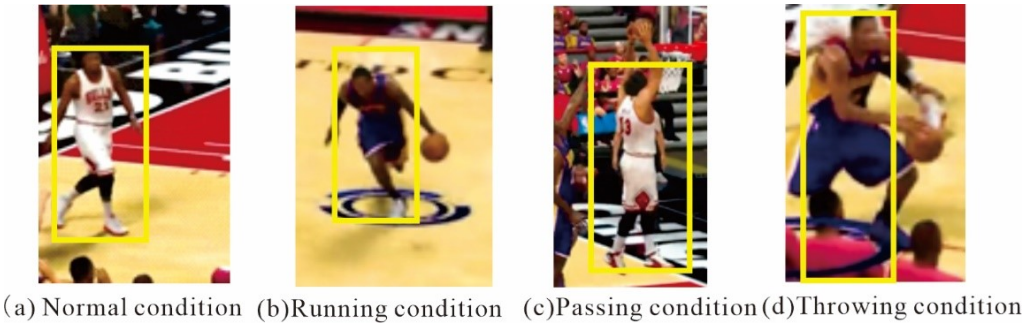
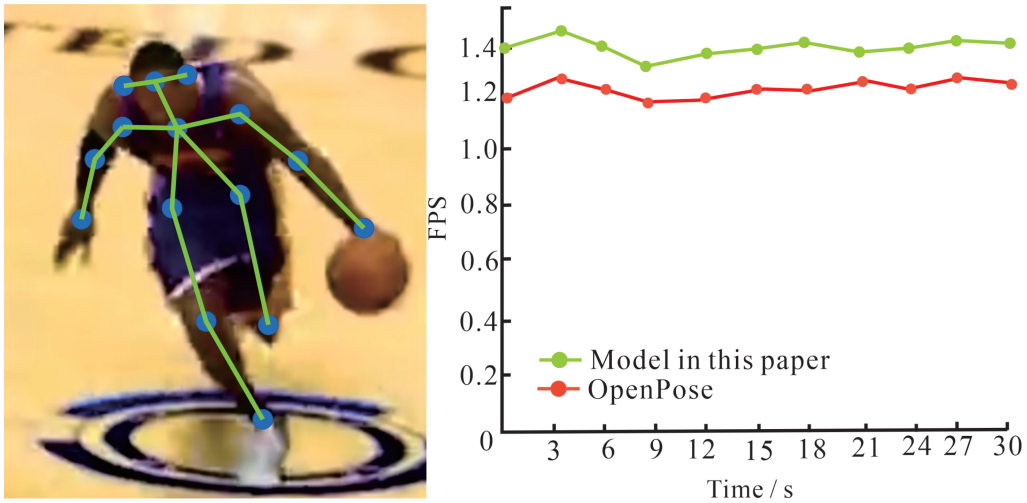


Figure 7. Detection of tracking frame monoblocks.

As shown in Figure 8 below, the location of the Tracking Framework is obtained, the centre point is selected as the location of the target, fed into the gesture estimation network for estimation, and the nodal data is output. In addition, for the later screen convolution action recognition, in order to make gesture estimation out of 200 frames the network runs first. In addition, the nodal point results are entered in the loop array (the loop array length is the space required for 200 frames). When the loop count is detected group is full of 200 frames, the system will take 200 frames of data from the latest video frames obtained from the loop array, and then reverse the sequence.The video frames are entered into the input cache of the network and inferred by the graph convolutional neural network, which eventually starts the graph convolutional network as a category. The new video frames are then overwritten with the oldest frames in the recurrent array, and then 200 frames are taken from the most recent video. Then the video frames are inverted and sent to the convolutional network for inference, and this is repeated until the video ends or tracking fails.

Figure 8. Overall detection of the tracking frame.

Figure 9 shows the application effect of the improved OpenPose model in a realistic environment. The prediction index of the improved lightweight OpenPose model decreases slightly compared with the original model, by about 3. However, the AP^L value of the lightweight model is not much different from the original model when detecting a larger size human body. This is because in this paper, the large size 7 × 7 convolutional kernel is replaced by a small size 3 × 3 convolutional kernel, and the small convolutional kernel can map the larger size object more nonlinearly in the feature extraction process, so it performs better in this aspect.



(a) Model actual application effect picture (b) Video frame rate comparison chart

Figure 9. Model application effect chart and video frame rate comparison chart.

In this paper, we compare the experimental results of GNNet with other advanced human pose estimation networks, and Table 2 shows the validation results of GNNet with other advanced networks on the COCO 2017 validation set. The proposed GNNet network structure is based on HRNet, and the GCTblock is used to model the channels more efficiently, thus enabling the Gateblock module and the Gateneck module extracts channel features more accurately, and improves the fusion module between different resolution subnets of each stage by using the Non-local block. This paper adds the Non-local block before the fusion of different resolution representations, which can extract more useful spatial feature information from different resolution representations, thus enabling better information fusion between multi-resolution representations.

Table 2. The results on COCO VAL 2017.

Method	Input Size	GFLOPS	AP^{50}	AP^{75}	mAP	AR
CPN [50]	256 * 192	6.20	-	-	68.6	-
CPN + OHKM [50]	256 * 192	6.20	-	-	69.4	-
HRNet-32 [51]	256 * 192	7.10	89.5	80.7	73.4	78.9
SimpleBaseline-50 [52]	256 * 192	8.90	88.6	78.3	70.4	76.3
SimpleBaseline-101 [52]	256 * 192	12.40	89.3	79.3	71.4	77.1
SimpleBaseline-152 [52]	256 * 192	15.70	89.5	79.8	72.0	77.8
Ours	256 * 192	8.1	**91.2**	**81.5**	**74.3**	**79.3**

This table shows the performance comparison of several different methods on the COCO VAL 2017 dataset for human pose estimation. The input size of all the methods is 256 * 192, and the GFLOPS (floating point operations per second) metric is used to measure the computational efficiency of each method. The evaluation metrics used are AP (average precision) at IoU thresholds of 0.5 and 0.75, mAP (mean average precision), and AR (average recall) at a fixed false positive rate.

Among the methods, CPN and CPN+OHKM do not have reported results for AP at either threshold, while HRNet-32, SimpleBaseline-50, SimpleBaseline-101, and SimpleBaseline-152 achieve AP values ranging from 88.6 percent to 89.5 percent at IoU threshold of 0.5, and from 78.3 percent to 79.8 percent at IoU threshold of 0.75. Our method outperforms all the other methods with an AP of 91.2 percent at IoU threshold of 0.5 and 81.5 percent at IoU threshold of 0.75, achieving the highest mAP of 74.3 percent and the highest AR of 79.3 percent. Compared with several other recent research methods, the method used in this paper has a better performance in AP50, AP75, mAP and AR metrics, which are generally better than other methods, and is also better represented in publicly available

datasets, which can show that the proposed method based on graph convolutional neural network and single body pose estimation has a better application in human pose estimation.

5. Discussion

In this paper, we present a rigorous analysis of existing models for human action recognition based on graphical convolutional networks. Specifically, we propose a novel approach that combines top-down single target pose estimation based on multi-branch self-calibration networks with graphical convolutional neural networks for human action recognition of skeletal data. This approach is designed to ensure adequate extraction of human pose features while reducing the time required for extraction, thereby improving the accuracy and efficiency of object detection and human pose estimation.

To evaluate the performance of our proposed approach, we conducted experiments on both a standard dataset and a custom basketball game dataset. The results show that our approach outperforms existing models, such as the spatiotemporal graph convolutional network and spatial transformation network, in terms of accuracy. Moreover, our approach significantly improves the performance on the custom basketball game dataset relative to existing methods.

Despite the success of our approach, we acknowledge that there are still some limitations and challenges that need to be addressed in future research. For instance, collecting "negative sample" data for special poses, such as falls and rollovers, is expensive but necessary in some fields. Additionally, labeling a 3D pose is more challenging than labeling a 2D pose. There is a need for further research to improve the sequence model of convolutional networks. To address these challenges, we propose that deep convolutional neural networks, such as the "Inception" structure, and extended convolutional networks, could accommodate growing datasets and reduce training time. Furthermore, Park's approach of learning the relative 3D positions between the joints of the human body through convolutional networks could improve the accuracy of 3D pose estimation from images.

In conclusion, this paper provides a rigorous analysis of existing models for human action recognition based on graphical convolutional networks and proposes a novel approach that combines top-down single-target pose estimation with graphical convolutional neural networks for human action recognition of skeletal data. Our approach shows promising results in improving the accuracy and efficiency of object detection and human pose estimation. However, future research is needed to address the challenges and limitations of human pose estimation, and we suggest deep convolutional neural networks and extended convolutional networks as potential solutions.

6. Conclusions

The field of sports artificial intelligence (AI) has received significant attention, with the importance of human pose estimation and its potential applications widely recognized. However, the research also reveals certain challenges and obstacles that must be addressed. To this end, this paper proposes a new approach that employs graph neural network combined with high resolution network (HRNET) as the base network model, introducing a linear transformation to generate redundant feature maps to reduce the network parametric size and network complexity. Additionally, an in-depth analysis of the GAFF module is incorporated into the feature attention fusion module to improve the feature fusion capability of the network. Experimental results on self-collected and MSCOCO2017 datasets demonstrate that the proposed approach reduces the network's covariance and complexity while improving its accuracy for human pose estimation.

However, the proposed method still has some shortcomings. For example, 3D pose estimation based on RGB video lacks auxiliary information such as depth information, making the calculation method less accurate. Additionally, the method of mapping 2D to 3D joint points is not yet perfect. While theoretically, 3D skeletal points contain richer inter-joint skeletal features for human action recognition, the study found that 2D joint points are more accurate than 3D skeletal points in action recognition. In future studies,

multiple extracted 2D skeleton points from pictures taken from different angles could be used to train the 2D to 3D skeleton mapping relationship.

The research on sports AI involves human wearable devices and computer vision technology analysis, including action recognition, training monitoring and feedback, game performance evaluation, and sports injury recovery. The theoretical aspect focuses on more efficient and convenient algorithms, with competitive sports as the main focus, particularly team sports such as soccer, supplemented by popular sports such as fitness. The research trend is toward combining algorithm theory with engineering for more practical applications. The future direction of sports AI research is to find more efficient and practical AI algorithms and models by continuously improving and matching, developing smarter wearable sensors, and enhancing the deep learning capability of visual analysis systems.

Author Contributions: C.D. conceived and initialized the research, conceived the algorithms, B.H. designed the experiments; W.L. evaluated the experiments, reviewed the paper, collected and analyzed the data; J.S. wrote the paper and organize article format. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Gomes, J.F.S.; Leta, F.R. Applications of computer vision techniques in the agriculture and food industry: A review. *Eur. Food Res. Technol.* **2012**, *235*, 989–1000. [CrossRef]
2. Song, Y.; Demirdjian, D.; Davis, R. Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2012**, *2*, 1–28. [CrossRef]
3. Shotton, J.; Girshick, R.; Fitzgibbon, A.; Sharp, T.; Cook, M.; Finocchio, M.; Moore, R.; Kohli, P.; Criminisi, A.; Kipman, A.; et al. Efficient human pose estimation from single depth images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 2821–2840. [CrossRef] [PubMed]
4. Fastovets, M.; Guillemaut, J.-Y.; Hilton, A. Athlete pose estimation by non-sequential key-frame propagation. In Proceedings of the 11th European Conference on Visual Media Production, London, UK, 13–14 November 2014; pp. 1–9.
5. Chun, S.; Ghalehjegh, N.H.; Choi, J.; Schwarz, C.; Gaspar, J.; McGehee, D.; Baek, S. Nads-net: A nimble architecture for driver and seat belt detection via convolutional neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
6. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
7. Yoon, Y.; Yu, J.; Jeon, M. Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition. *Appl. Intell.* **2022**, *52*, 2317–2331. [CrossRef]
8. Simonovsky, M.; Komodakis, N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017 ; pp. 3693–3702.
9. Maxwell, J.A.; Mittapalli, K. Realism as a stance for mixed methods research. In *SAGE Handbook of Mixed Methods in Social & Behavioral Research*; Sage: Thousand Oaks, CA, USA , 2010; pp. 145–168.
10. Bouraffa, T.; Feng, Z.; Yan, L.; Xia, Y.; Xiao, B. Multi–feature fusion tracking algorithm based on peak–context learning. *Image Vis. Comput.* **2022**, *123*, 104468. [CrossRef]
11. Gamboa, H.; Fred, A. A behavioral biometric system based on human-computer interaction. In *Biometric Technology for Human Identification*; SPIE: Bellingham, WA, USA, 2004; Volume 5404, pp. 381–392.
12. Wu, S.; Wang, J.; Ping, Y.; Zhang, X. Research on individual recognition and matching of whale and dolphin based on efficientnet model. In Proceedings of the 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 15–17 July 2022; pp. 635–638.
13. Zhang, X.; Ping, Y.; Li, C. Artificial intelligence-based early warning method for abnormal operation and maintenance data of medical and health equipment. In Proceedings of the IoT and Big Data Technologies for Health Care: Third EAI International Conference, IoTCare 2022, Virtual, 12–13 December 2022; Springer: Berlin/Heidelberg, Germany, 2023; pp. 309–321.
14. Farin, D.; Krabbe, S.; de With, P.H.N.; Effelsberg, W. Robust camera calibration for sport videos using court models. In *Storage and Retrieval Methods and Applications for Multimedia 2004*; SPIE: Bellingham, WA, USA, 2003; Volume 5307, pp. 80–91.

15. Dargan, S.; Kumar, M. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Syst. Appl.* **2020**, *143*, 113114. [CrossRef]
16. Roussaki, I.; Strimpakou, M.; Kalatzis, N.; Anagnostou, M.; Pils, C. Hybrid context modeling: A location-based scheme using ontologies. In Proceedings of the Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOMW'06), Pisa, Italy, 13–17 March 2006; p. 6.
17. Albert, J. Baseball data at season, play-by-play, and pitch-by-pitch levels. *J. Stat. Educ.* **2010**, *18* . [CrossRef]
18. Doroniewicz, I.; Ledwoń, D.J.; Affanasowicz, A.; Kieszczyńska, K.; Latos, D.; Matyja, M.; Mitas, A.W.; Myśliwiec, A. Writhing movement detection in newborns on the second and third day of life using pose-based feature machine learning classification. *Sensors* **2020**, *20*, 5986. [CrossRef]
19. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
20. Tompson, J.J.; Jain, A.; LeCun, Y.; Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
21. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human pose estimation with iterative error feedback. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016 ; pp. 4733–4742.
22. Farrukh, W.; Haar, D.v.d. Computer-assisted self-training for kyudo posture rectification using computer vision methods. In Proceedings of the Fifth International Congress on Information and Communication Technology, London, UK, 17–19 December 2021 ; Springer: Berlin/Heidelberg, Germany, 2021; pp. 202–213.
23. Fan, X.; Zhao, S.; Zhang, X.; Meng, L. The impact of improving employee psychological empowerment and job performance based on deep learning and artificial intelligence. *J. Organ. End User Comput. (JOEUC)* **2023**, *35*, 1–14. [CrossRef]
24. Paul, M.K.A.; Kavitha, J.; Rani, P.A.J. Key-frame extraction techniques: a review. *Recent Patents Comput. Sci.* **2018**, *11*, 3–16. [CrossRef]
25. Cabán, C.C.T.; Yang, M.; Lai, C.; Yang, L.; Subach, F.V.; Smith, B.O.; Piatkevich, K.D.; Boyden, E.S. Tuning the sensitivity of genetically encoded fluorescent potassium indicators through structure-guided and genome mining strategies. *ACS Sens.* **2022**, *7*, 1336. [CrossRef] [PubMed]
26. Li, C.; Chen, Z.; Jiao, Y. Vibration and bandgap behavior of sandwich pyramid lattice core plate with resonant rings. *Materials* **2023**, *16*, 2730. [CrossRef]
27. Nasr, M.; Ayman, H.; Ebrahim, N.; Osama, R.; Mosaad, N.; Mounir, A. Realtime multi-person 2d pose estimation. *Int. J. Adv. Netw. Appl.* **2020**, *11*, 4501–4508. [CrossRef]
28. Osokin, D. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv* **2018**, arXiv:1811.12004.
29. Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
30. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019 ; Springer: Berlin/Heidelberg, Germany, 2016; pp. 483–499.
31. Presti, L.L.; Cascia, M.L. 3d skeleton-based human action classification: A survey. *Pattern Recognit.* **2016**, *53*, 130–147. [CrossRef]
32. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7753–7762.
33. Gärtner, E.; Pirinen, A.; Sminchisescu, C. Deep reinforcement learning for active human pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10835–10844.
34. Vila, M.; Bardera, A.; Xu, Q.; Feixas, M.; Sbert, M. Tsallis entropy-based information measures for shot boundary detection and keyframe selection. *Signal Image Video Process.* **2013**, *7*, 507–520. [CrossRef]
35. Jain, A.K. Data clustering: 50 years beyond k-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
36. Hara, K.; Chellappa, R. Growing regression tree forests by classification for continuous object pose estimation. *Int. J. Comput. Vis.* **2017**, *122*, 292–312. [CrossRef]
37. Papadaki, S.; Wang, X.; Wang, Y.; Zhang, H.; Jia, S.; Liu, S.; Yang, M.; Zhang, D.; Jia, J.-M.; Köster, R.W.; et al. Dual-expression system for blue fluorescent protein optimization. *Sci. Rep.* **2022**, *12*, 1–16. [CrossRef] [PubMed]
38. Ning, X.; Gong, K.; Li, W.; Zhang, L.; Bai, X.; Tian, S. Feature refinement and filter network for person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 3391–3402. [CrossRef]
39. Ning, X.; Nan, F.; Xu, S.; Yu, L.; Zhang, L. Multi-view frontal face image generation: A survey. *Concurr. Comput. Pract. Exp.* **2020** , e6147. [CrossRef]
40. Ning, X.; Duan, P.; Li, W.; Zhang, S. Real-time 3d face alignment using an encoder-decoder network with an efficient deconvolution layer. *IEEE Signal Process. Lett.* **2020**, *27*, 1944–1948. [CrossRef]
41. He, F.; Ye, Q. A bearing fault diagnosis method based on wavelet packet transform and convolutional neural network optimized by simulated annealing algorithm. *Sensors* **2022**, *22*, 1410. [CrossRef] [PubMed]
42. Chen, C.-C.; Chang, C.; Lin, C.-S.; Chen, C.-H.; Chen, I.C. Video based basketball shooting prediction and pose suggestion system. *Multimed. Tools Appl.* **2023**, 1–20. [CrossRef]

43. Zhang, Y.-H.; Wen, C.; Zhang, M.; Xie, K.; He, J.-B. Fast 3d visualization of massive geological data based on clustering index fusion. *IEEE Access* **2022**, *10*, 28821–28831. [CrossRef]
44. Zhang, M.; Xie, K.; Zhang, Y.-H.; Wen, C.; He, J.-B. Fine segmentation on faces with masks based on a multistep iterative segmentation algorithm. *IEEE Access* **2022**, *10*, 75742–75753. [CrossRef]
45. Saiki, Y.; Kabata, T.; Ojima, T.; Kajino, Y.; Inoue, D.; Ohmori, T.; Yoshitani, J.; Ueno, T.; Yamamuro, Y.; Taninaka, A.; et al. Reliability and validity of openpose for measuring hip-knee-ankle angle in patients with knee osteoarthritis. *Sci. Rep.* **2023**, *13*, 3297. [CrossRef]
46. Hooren, B.V.; Pecasse, N.; Meijer, K.; Essers, J.M.N. The accuracy of markerless motion capture combined with computer vision techniques for measuring running kinematics. *Scand. J. Med. Sci. Sport.* **2023**, *33*, 966–978.
47. Yi, G.; Wu, H.; Wu, X.; Li, Z.; Zhao, X. Human action recognition based on skeleton features. *Comput. Sci. Inf. Syst.* **2023**, *20*, 537–550. [CrossRef]
48. Gao, M.; Li, J.; Zhou, D.; Zhi, Y.; Zhang, M.; Li, B. Fall detection based on openpose and mobilenetv2 network. *IET Image Process.* **2023**, *17*, 722–732. [CrossRef]
49. Dewi, C.; Chen, A.P.S.; Christanto, H.J. Deep learning for highly accurate hand recognition based on yolov7 model. *Big Data Cogn. Comput.* **2023**, *7*, 53. [CrossRef]
50. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
51. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
52. 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, Venice, Italy, 22–29 October 2017; pp. 2640–2649.