electronics

MDPI

Article

Human Action Recognition Based on Skeleton Information and Multi-Feature Fusion

Li Wang [1,2], Bo Su [1,3], Qunpo Liu [1,3], Ruxin Gao [1,3], Jianjun Zhang [1,3] and Guodong Wang [4,*]

[1] School of Electrical Engineering & Automation, Henan Polytechnic University, Jiaozuo 454003, China
[2] Henan Key Laboratory of Intelligent Detection and Control of Coal Mine Equipment, Jiaozuo 454003, China
[3] Henan International Joint Laboratory of Direct Drive and Control of Intelligent Equipment, Jiaozuo 454003, China
[4] Computer Science Department, Massachusetts College of Liberal Arts, North Adams, MA 01247, USA
* Correspondence: guodong.wang@mcla.edu

Abstract: Action assessment and feedback can effectively assist fitness practitioners in improving exercise benefits. In this paper, we address key challenges in human action recognition and assessment by proposing innovative methods that enhance performance while reducing computational complexity. Firstly, we present Oct-MobileNet, a lightweight backbone network, to overcome the limitations of the traditional OpenPose algorithm's VGG19 network, which exhibits a large parameter size and high device requirements. Oct-MobileNet employs octave convolution and attention mechanisms to improve the extraction of high-frequency features from the human body contour, resulting in enhanced accuracy with reduced model computational burden. Furthermore, we introduce a novel approach for action recognition that combines skeleton-based information and multiple feature fusion. By extracting spatial geometric and temporal characteristics from actions, we employ a sliding window algorithm to integrate these features. Experimental results show the effectiveness of our approach, demonstrating its ability to accurately recognize and classify various human actions. Additionally, we address the evaluation of traditional fitness exercises, specifically focusing on the BaDunJin movements. We propose a multimodal information-based assessment method that combines pose detection and keypoint analysis. Label sequences are obtained through a pose detector and each frame's keypoint coordinates are represented as pose vectors. Leveraging multimodal information, including label sequences and pose vectors, we explore action similarity and perform quantitative evaluations to help exercisers assess the quality of their exercise performance.

Keywords: motion recognition; backbone network; motion evaluation

Citation: Wang, L.; Su, B.; Liu, Q.; Gao, R.; Zhang, J.; Wang, G. Human Action Recognition Based on Skeleton Information and Multi-Feature Fusion. *Electronics* **2023**, *12*, 3702. https://doi.org/10.3390/electronics12173702

Academic Editors: Rania Hodhod and Mohammad Jafari

Received: 3 August 2023
Revised: 28 August 2023
Accepted: 28 August 2023
Published: 1 September 2023

1. Introduction

Exercise is essential in daily life. With the acceleration of the pace of life, people tend to exercise at home and evaluate their exercise effectiveness by themselves. However, ordinary self-study methods lack professional guidance, and incorrect or improper body movements can lead to a decrease in exercise effectiveness and even cause physical harm. Therefore, it is necessary to analyze and evaluate exercise and provide feedback for achieving better results.

Currently, most exercise analysis relies on specialized sensor devices [1]. For example, Albert et al. [2] developed a home exercise system based on inertial measurement components, which allows users to practice kicking exercises at home by wearing sensors. Gupta et al. [3] designed a yoga-assisted exercise system that helps amateur enthusiasts learn correct yoga movements without the supervision of a coach. The motion sensors used in this system mainly include accelerometers and gyroscopes. Although these wearable sensors have good data acquisition and analysis capabilities, they require exercisers to wear related devices, which causes inconvenience to exercise and also has certain invasiveness and safety risks to the human body.

With the development and popularization of computers and image technology, visual-based action recognition has gradually attracted people's attention [4]. The idea of visual-based recognition is that there is action information hidden in human skeletal data, and action recognition can be achieved through methods such as feature extraction and information mining. For example, the authors of [5] proposed a machine learning-based shooting action recognition method. This method can extract multidimensional motion posture features and use those features to recognize actions. In addition, Xu et al. [6] designed an end-to-end network to improve the efficiency of optical flow feature extraction, and combined spatiotemporal features to achieve human action recognition. These methods have brought convenience and achieved certain performance. However, it requires a large amount of annotated data and a long training time if directly extracting video features based on deep learning.

In this paper, we propose a visual-based motion recognition algorithm and apply it to the fitness exercise "BaDuanJin", an exercise that focuses on a mind–body integration [7]. The proposed method does not require a large amount of annotated data and it has low computational complexity, making it suitable for motion videos with varying lengths. In addition, we propose a feedback method for trainers to evaluate their training effectiveness and obtain recommendations accordingly. In this feedback method, label sequence and pose vector are used to evaluate the video motion similarity, which assists practitioners in correcting their actions and obtaining suggestions.

The main contribution of this paper includes the following.

- This paper proposes a lightweight backbone network called Oct-MobileNet to address the problem of high parameter volume in the traditional OpenPose algorithm's VGG19 backbone network. The proposed method utilizes the attention mechanism and octave convolution to enhance the network's ability to extract high-frequency features of human body contours while reducing the model's computational complexity.
- A motion recognition method based on skeleton information and multi-feature fusion is proposed in this paper. Multiple features are extracted from the spatial geometry and temporal characteristics, and a sliding window algorithm is used to fuse them.
- A multimodal information-based motion evaluation method is proposed for the fitness exercise. The method uses a pose detector to obtain the label sequence of standard motions and represents the coordinates of each frame's key points as a pose vector. The method extracts motion similarity from multiple modal information, such as label sequence and pose vector, and performs a quantitative evaluation for exercisers.

The rest of the paper is organized as follows. In Section 2, the related work is introduced. Section 3 introduces the Oct-OpenPose and Coordinates Preprocess. In Section 4, we elaborate on the Action Recognition Model. Section 5 presents the Evaluation Methods for BaDuanJin Movements. Concluding remarks are drawn in Section 6.

2. Related Work

Human action recognition is one of the fundamental ways to achieve human–computer interaction aiming at describing human behavior in images or videos. From the perspective of information acquisition, the methods can be broadly classified into two categories [8]: sensor-based action recognition and vision-based action recognition.

Sensor-based action recognition has emerged as a promising field of research with numerous applications in fields such as healthcare, sports analysis, and human–computer interactions [9]. In particular, wearable sensors are placed on important joint positions of the human body to collect motion data in real time and transmit it to the upper computer. The upper computer then analyzes and processes the data, thus achieving human motion recognition. Pansiot et al. [10] designed a head-mounted inertial sensor for swimming motion analysis. By recording features such as pitch angle and rotation angle, they analyzed the movement details of swimmers and developed a swimming monitoring system for training guidance. Ma et al. [11] designed a wireless network module based on MEMS inertial sensors [12] and Zigbee [13] used the SVM classification algorithm [14] for human

motion recognition. This work showed desirable performance in detecting abnormal behaviors such as falls. Authors of [15] developed a wearable system using an accelerometer, an analog-to-digital converter, and a WiFi module to acquire and transmit human motion data. They used the KNN algorithm [16] to recognize four types of daily behaviors: lying down, sitting, standing, and walking, achieving an accuracy of 93%. Authors of [17] used a smartphone on the waist of subjects to measure linear acceleration and angular velocity along three axes using the built-in accelerometer and gyroscope of the phone. The LightGBM [18] algorithm was used to recognize and classify six types of movements, including going up and down stairs, sitting, and standing.

While sensor-based action recognition has shown promising results in various applications, it also has limitations that need to be considered. The biggest limitation is that it relies on the availability and proper placement of sensors on the body to capture motion data. Compared to sensor-based action recognition, vision-based action recognition does not use any extra sensors, which makes it a preferred choice in many scenarios. Recently, there has been some research in this field. For example, Sun et al. [19] proposed a method for extracting key frames from video streams, and combined it with a posture estimation algorithm to extract skeleton information for golf action comparison recognition.

In addition to action recognition, human action evaluation is also an important branch of motion analysis. By measuring the difference or similarity between two actions, the consistency of the practice action and the standard action can be judged. For example, Xu et al. [20] designed two complementary long short-term memory networks to extract information from figure skating videos and used aggregate feature information to predict scores. The proposed model achieved desirable performance on their self-built data set. Another work is provided by authors of [21], who proposed a method for evaluating the similarity of etiquette actions. First, the video keyframes were extracted using the frame difference method, and then the evaluation criteria were set based on standard actions. Action scoring was carried out by calculating the similarity of limb vector angles and foot distance in keyframes. The above-mentioned methods generally require a substantial quantity of annotated data and lengthy training times, which results in relatively limited model generalization capabilities.

Based on the above discussion, it is evident that sensor-based action assessment approaches heavily depend on hardware devices, which presents challenges in terms of widespread adoption. On the other hand, deep learning methods that directly extract video features necessitate a substantial volume of annotated data and entail lengthy training durations. Therefore, there is a need for an alternative approach that eliminates the requirement for extensive annotated data, possesses low computational complexity, and proves suitable for fitness routines of varying durations.

3. Oct-OpenPose and Coordinates Preprocess

In this section, we analyze the current popular human pose estimation methods and then propose a method that meets the requirements of body movement recognition discussed in the related work section. In particular, we focus on fitness exercises in normal scenarios, and a lightweight model is needed for human pose estimation.

Visual-based human pose estimation methods can be divided into two categories: the top-down approach and the bottom-up approach. The top-down approach detects each person in the image first and then performs key point detection for each individual. The typical top-down approach algorithms include RMPE [22] and CPN [23]. In contrast, the bottom-up approach detects all the key points of all people in the image first and then assigns key points to each individual. The typical bottom-up approach algorithms include OpenPose [24] and HigherHRNet [25].

The bottom-up approach is relatively faster and more robust, so in this paper, we use the bottom-up approach for keypoint detection and multi-body pose estimations. We have two options: HigherHRNet and OpenPose. The HigherHRNet network is complex and usually focuses on solving the pose estimation problem for dense, small-sized people,

which does not meet the scenario of fitness exercise. Therefore, we choose OpenPose in this paper. The traditional OpenPose uses VGG19 as the backbone network to extract features from input images. The VGG19 network, due to its extensive parameter count, is not considered a lightweight backbone network. Consequently, the traditional OpenPose model utilizing VGG19 fails to fulfill the real-time performance demanded by the scenario discussed in this paper.

Natural images can be decomposed into low-frequency components and high-frequency components. Octave convolution divides the convolutional feature map into low-frequency components and high-frequency components. Information interactions between the two frequency groups are achieved through up-sampling and down-sampling, and ultimately, the original feature map can be reassembled. We can achieve the following advantages by replacing traditional convolutions with Octave convolutions. (1) Reduced storage and computational load. The size of the low-frequency feature map can be halved, effectively reducing storage and computational resources. This enhancement contributes to improving the speed of detecting human key points. (2) Enhanced receptive field. With the feature map size reduced while keeping the convolutional kernel size constant, the receptive field increases. This expansion enables better capture of contextual information in scenarios like the "BaDuanJin" movement, ultimately enhancing recognition performance. (3) Preserved high-frequency information. Octave convolution fully retains high-frequency information, and since the human body outline constitutes critical high-frequency information for action recognition, this preservation is important.

Based on the structure design of OpenPose and the above discussion, we propose an improved model named Oct-OpenPose in this paper. The network structure of the Oct-OpenPose is shown in Figure 1. The Predicted Affinity Fields (PAFs) prediction part of the network is reduced from the original four stages to three stages, and the Heatmaps prediction part is reduced to one stage. The first convolution kernel size of the three inside concatenated convolutions is changed to 1×1. The final convolution kernel is designed as a dilated convolution with a dilation factor of two to reduce the redundancy of the network. The new backbone network Oct-MobileNet using the Oct-OpenPose has greatly reduced the parameter quantity of the MobileNet. It also reduces low-frequency information redundancy and focuses more on high-frequency information so that the extracted deep features contain more effective information.

In order to evaluate the performance of the proposed model, we used the Common Objects in the Context (COCO2017) data set to test the MobileNet and Oct-MobileNet networks. Specifically, we used the adaptive momentum method to train and optimize the models. The training was conducted for 3×10^5 iterations, with an initial learning rate of 4×10^{-4}, and a batch size of 24. Figure 2 compares the feature visualization results. It can be seen that the Oct-MobileNet places more emphasis on high-frequency feature components such as the human body contour when extracting deep features and suppresses other low-frequency components, which makes the PAFs of human key points more clear.

We also compared the accuracy of the proposed model as well as the network in Table 1. As can be seen from Table 1, the number of stages for PAFs and Heatmaps has been reduced from 4 + 2 to 3 + 1, with negligible loss in accuracy. When using a regular MobileNet as the backbone network, the model's accuracy decreased by 7.1% compared to the original network. However, Oct-MobileNet, which integrates the improved Octave convolution, places more emphasis on high-frequency features such as the human body outline and suppresses other low-frequency features when extracting features, resulting in a 6% increase in model accuracy compared to regular MobileNet. The accuracy of the improved model is only reduced by 1.2% compared to the original model. The detection speed of the improved model can reach 31 fps, which is 300% faster than the original model.

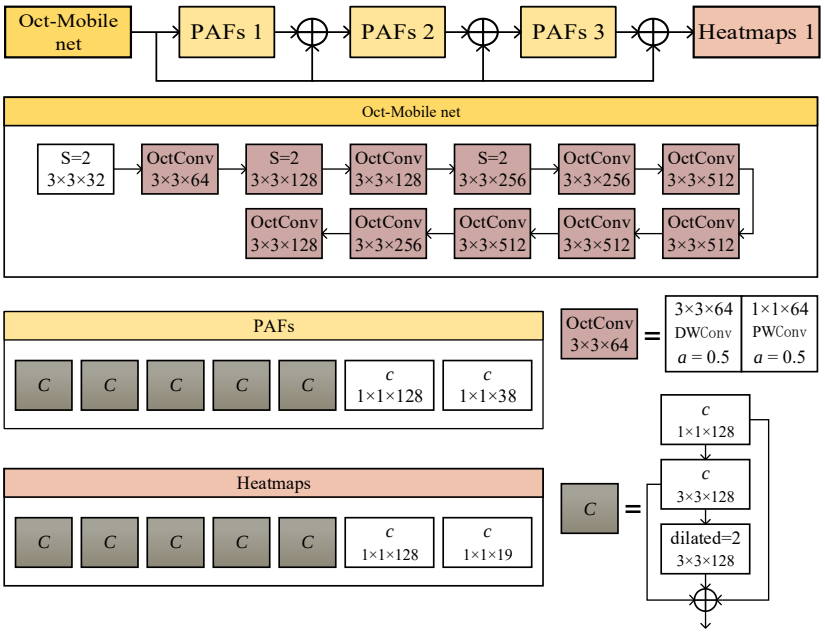


Figure 1. Architecture of the Oct-OpenPose.

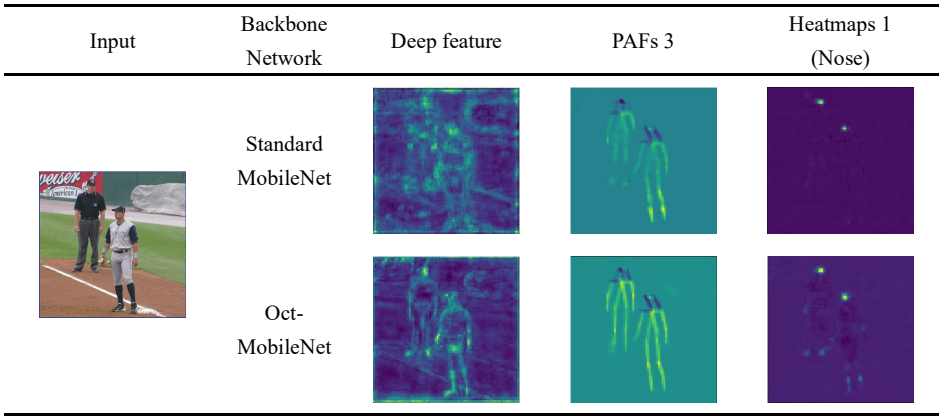


Figure 2. Feature visualization comparison results.

Table 1. Comparison of OpenPose models with different structures.

Models	Backbone	PAFs Stages	Heatmaps Stages	mAP (%)	Speed (fps)
OpenPose	VGG19	4	2	65.2	7.1
		3	1	64.6	8
Revised-OpenPose	MobileNet	3	1	57.5	29
	Oct-MobileNet	3	1	63.4	31

Besides feature extraction, we also normalize the coordinates of key points. The coordinates are normalized in order to solve the problem of non-uniformity caused by image size or lens distance variation. First, the input image is scaled so that an image with width and height (w, h) is first scaled to (1, h/w). The center position (X_c, Y_c) of the human body is then calculated based on the coordinates of 18 points according to Equation (1). Second, we calculate the height H between the neck and the hip. Note that the height H does not change for different actions. Finally, each keypoint is subtracted from

the central coordinate and divided by H according to Equation (2) to obtain the normalized coordinates (Xnew, Ynew), as shown in Figure 3.

$$X_c = \frac{\sum_{i=1}^{18} x_i}{18}, Y_c = \frac{\sum_{i=1}^{18} y_i}{18} \tag{1}$$

$$X_{new} = \frac{x_i - X_c}{H}, Y_{new} = \frac{y_i - Y_c}{H}, \tag{2}$$

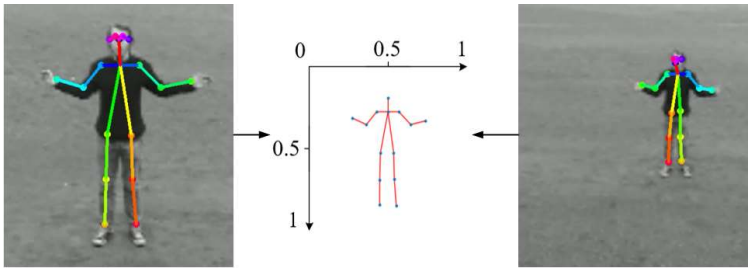


Figure 3. Skeleton data preprocessing.

4. Action Recognition Model

During human movement, the main movements are in the limbs and torso, which are not related to the face. Therefore, eye and ear coordinates are excluded, and the position information of the head is abstracted using the nose to represent the head position, reducing the total number of key points from 18 to 14. The diagram of geometric features is depicted in Figure 4, where the numbers represent the key joints of human body.

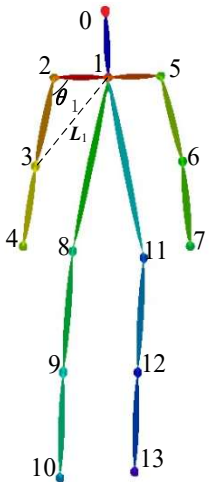


Figure 4. Diagram of geometric features of human body.

The spatial position of the human skeleton and its geometric relationship can be used to model human motion. In order for the classifier to better distinguish between different movements in the BaDuanJin, spatial geometry and temporal motion features are extracted from the skeletal coordinate information. Taking $L1$ and $\theta 1$ as an example, Table 2 shows that d(1–3) represents the pixel distance between the neck and right elbow, while $\angle(1-2-3)$ represents the angle formed by the neck, right shoulder, and right elbow, as shown in Figure 4. The calculation formulas are shown in Equations (3) and (4), where (x_1,y_1) and (x_3,y_3) represent the coordinates of key points 1 and 3, and a, b, and c represent the three sides of the triangle formed by the neck, right shoulder, and right elbow, respectively.

Table 2. Spatial geometric feature table.

Joint Distance	Symbol	Joint Angle	Symbol
$L1$	d(1–3)	$\theta 1$	$\angle(1-2-3)$
$L2$	d(2–4)	$\theta 2$	$\angle(2-3-4)$
$L3$	d(1–6)	$\theta 3$	$\angle(1-5-6)$
$L4$	d(5–7)	$\theta 4$	$\angle(5-6-7)$
$L5$	d(1–9)	$\theta 5$	$\angle(1-8-9)$
$L6$	d(8–10)	$\theta 6$	$\angle(8-9-10)$
$L7$	d(1–12)	$\theta 7$	$\angle(1-11-12)$
$L8$	d(11–13)	$\theta 8$	$\angle(11-12-13)$
$L9$	d(4–8)	$\theta 9$	$\angle(2-1-8)$
$L10$	d(7–11)	$\theta 10$	$\angle(5-1-11)$
$L11$	d(8–13)	$\theta 11$	$\angle(8-9-13)$
$L12$	d(10–11)	$\theta 12$	$\angle(11-12-10)$

$$L_1 = d_{(1-3)} = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} \tag{3}$$

$$\theta_1 = arccos(\frac{a^2 + c^2 - b^2}{2ac}) = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} \tag{4}$$

The motion feature reflects the changes in the human body over consecutive moments, such as the swing speed of certain joints or limbs. The calculation method for key point velocity is to divide the displacement into the x and y directions between adjacent frames T_k and T_{k+1} by the interval time t, as shown in Equation (5). The value of t is related to the video frame rate F. For example, when F is 25 fps, t = 0.04 s (1/25 = 0.04 s).

$$V_x = \frac{X_{T_{k+1}} - X_{T_k}}{t}, V_y = \frac{Y_{T_{k+1}} - Y_{T_k}}{t} \tag{5}$$

To achieve sequence-based feature extraction, new features need to be constantly extracted from the video sequence and old features need to be removed in real time. We use a sliding window algorithm to deal with the feature extraction, as shown in Figure 5. A window of size N is used to store data based on the time series. As time progresses, the window moves directionally, with new data added to the head of the window and old data pushed out from the tail. This process continues until the window has traversed through all the data. In this paper, the sliding window size N is set to 10 frames, and spatial geometry features and temporal motion features are calculated accordingly. The final aggregated feature vector is shown in Table 3, where the dimension of the fused feature reaches 772. In order to reduce the computational complexity during model training, the classical feature extraction and data representation technique Principal Component Analysis (PCA) [26] is used to compress the features to 100 dimensions.

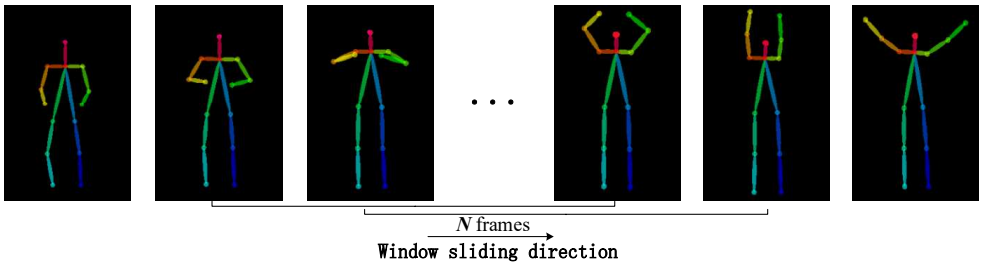


Figure 5. The process of sliding window algorithm.

Table 3. Feature List based on Video Sequences.

Feature Category	Feature Name	Dimension	Descriptions
Spatial Geometric Feature	Key points Position	280	Normalized coordinates of 14 key points
	Distance between key points	120	Euclidean distance between key points
	Joint angle	120	Angle formed by joints
Temporal Motion Feature	Key point velocity	252	Velocity of each key point in the x and y direction
Fused Feature	Spatial geometric + temporal motion	772	Combination of the above features

In this paper, since the scenario is action recognition based on time series, an LSTM-based classifier is proposed to deal with the classification. In particular, the input to the model is the 100-dimensional features that are aggregated and reduced by the sliding window with a time step T of 30. Two layers of LSTM units are designed in the LSTM block, and each layer has 128 neurons. After that, a fully connected layer (FC) with a size of 64 is used to further integrate the features. Finally, the softmax activation function is used to classify the movements.

Figure 6 shows the classification results and confusion matrix of the recognition results. It can be seen that the average recognition accuracy of each movement reaches 95.7%, and the frame rate of video detection can be maintained above 25 fps.

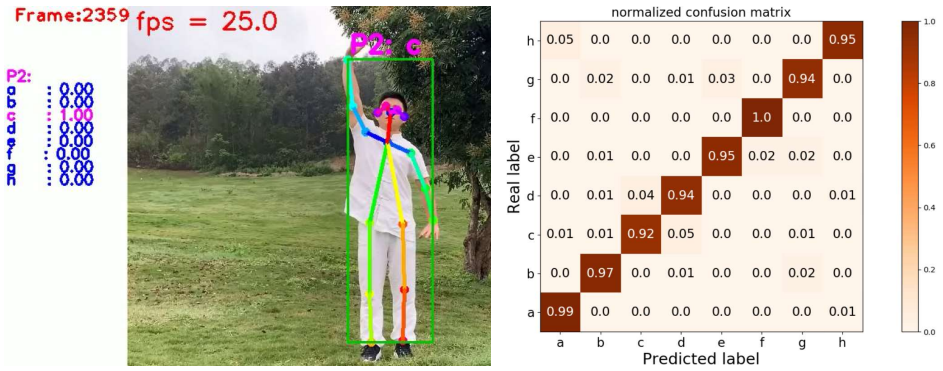


Figure 6. Classification results and confusion matrix.

5. The Evaluation Methods for BaDuanJin Movements

In the context of BaDuanJin movements, the mere recognition of the actions holds limited significance. The primary objective is to thoroughly analyze the disparities between the practitioner's movement sequences and the standard ones, facilitating a quantitative evaluation. This evaluation aims to assist the practitioner in rectifying individual movements, ultimately enhancing the overall effectiveness and benefits of the exercise routine. This section proposes a multi-modal information-based movement evaluation method, as shown in Figure 7. The first modality information is joint vector information extracted from each frame of the video, and the overall similarity between the test movement and the standard movement is measured by evaluating the similarity of joint vectors. The second modality information is the action label detected by the pose detection model for each frame, and the periodicity and synchronization of the two movements are evaluated by analyzing the similarity of the label sequence.

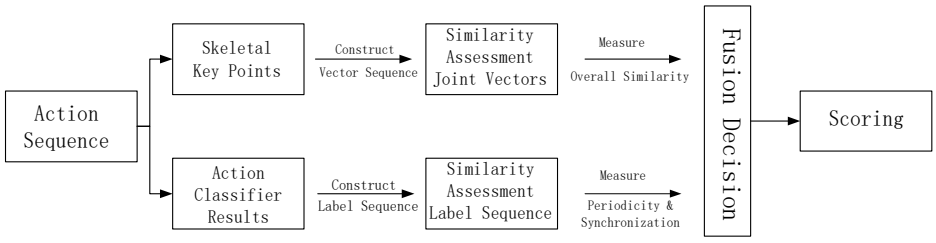


Figure 7. Process of multi-modal information-based movement evaluation.

5.1. Similarity Evaluation Based on Joint Vectors

By improving OpenPose to detect BaDuanJin movements in videos, the coordinates of key points for each frame of the human body can be extracted. These coordinates can be further organized into vectors to characterize the movement of the human body in each frame, and similarity can be analyzed through the cosine distance. At the same time, the length of the BaDuanJin movement sequence varies, making it suitable for dynamic time warping (DTW) algorithm processing. Inspired by this, this section proposes a similarity evaluation method based on joint vectors that combines cosine similarity and the DTW algorithm to calculate the similarity of movement sequences.

Let A = $(a_1, a_2, ..., a_n)$ and B = $(b_1, b_2, ..., b_m)$ be two time series with lengths n and m, respectively, where $n \neq m$. To align the two sequences on the time axis, a matrix grid with m rows and n columns is constructed as shown in Figure 8. The value in the i-th row and j-th column of the grid represents the distance between b_i and a_j, denoted as $d(i, j)$. Then, the DTW distance calculation method for the time series A and B is shown in Equation (6). Equation (6) can characterize the similarity between the two sequences, where a smaller distance indicates a higher similarity.

$$D(i, j) = d(i, j) + min[D(i-1, j), D(i, j-1), D(i-1, j-1)] \quad (6)$$

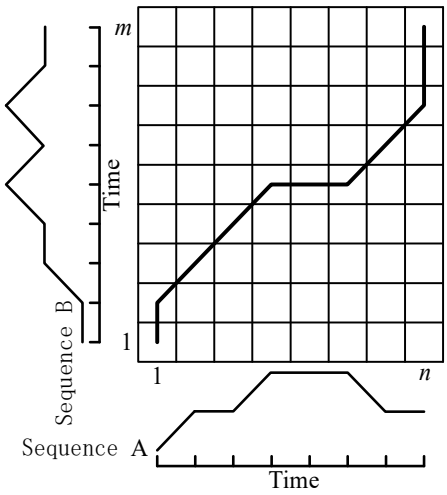


Figure 8. Schematic diagram of the sequence alignment based on DTW algorithm.

In order to facilitate analysis and calculation, a minimum bounding box is created around the human in the image using the coordinates (x_i, y_i) of the 18 key points outputted by the model, where $i \in [0, 17]$. Then, we create a new coordinate system, and the position of the key points in the new coordinate system is calculated according to Equation (7), where x_{min} and y_{max} represent the position of the lower-left endpoint of the bounding box.

$$x_{inew} = x_i - x_{min}; y_{inew} = y_{max} - y_i; i \in [0, 17] \quad (7)$$

For the movement A in a certain frame of the video, the preprocessed key point coordinates are represented as a high-dimensional vector in order, i.e., A = $[a_0, a_2, ..., a_{13}]$, where $a_0 \sim a_{13}$ are the position coordinates of 14 key points. Then, the similarity between two frames of movements A and B can be transformed into the similarity between vectors A and B. Suppose that the number of frames in video M and N is m and n, respectively, during a time interval t, and the human body movements in the two videos can be characterized as sequences $M = [m_1, m_2, ..., m_i, ..., m_m]$ and $N = [n_1, n_2, ..., n_j, ..., n_n]$, where m_i and n_j are

vectors composed of key point coordinates in a certain frame. By substituting Equation (8) into Equation (6), the distance between the two sequences can be calculated.

$$d(i,j) = 1 - cos(m_i, n_j) \tag{8}$$

$$Similarity\ Score = (1 - \frac{D}{D_{max}}) \times 100 \tag{9}$$

The $d(i,j)$ in Equation (8) represents the cosine distance between m_i and n_j, and the similarity score between the two sequences is finally calculated using Equation (9), which represents the overall similarity between the two sequences. D represents the DTW distance $D(m,n)$ between the sequences M and N. D_{max} is the distance when the sequence similarity reaches its minimum, and at this time, the cosine distance of each element reaches its maximum. The pseudo-code of the algorithm is given in Algorithm 1.

Algorithm 1 Similarity evaluation based on joint vectors

Input:
  Template Sequence $M = \{m_1, m_2, ..., m_i, ..., m_m\}$
  State Test sequence $N = \{n_1, n_2, ..., n_j, ..., n_n\}$
Output:
  Similarity Score
$D(0,0) \leftarrow 0$
for $(i = 1; i < m; i++)$ **do**
  for $(j = 1; i < n; j++)$ **do**
    $d(i,j) \leftarrow 1 - \cos(m_i, n_j)$
    $D(i,j) \leftarrow d(i,j) + min(D(i-1,j), D(i,j-1), D(i-1,j-1))$
  end for
end for
Calculate D_{max} for template sequence M
Similarity Score $\leftarrow (1 - D(m,n)/D_{max}) \times 100$

5.2. Similarity Evaluation Based on Label Sequence

The BaDuanJin movements exhibit robust regularity, characterized by their periodic nature. Each movement follows a specific pattern, commencing from the preparatory position, progressing through several essential postures, and concluding by returning to the preparatory position. For example, as depicted in Figure 9, movement b is decomposed into posture combinations of 1-4-5-1-4-6-1, and a complete cycle is repeated three times. Similarly, other movements can also be decomposed in a similar way. The eight movements ($a \sim h$) are divided into 21 key postures, which will be used as labels for similarity evaluation.

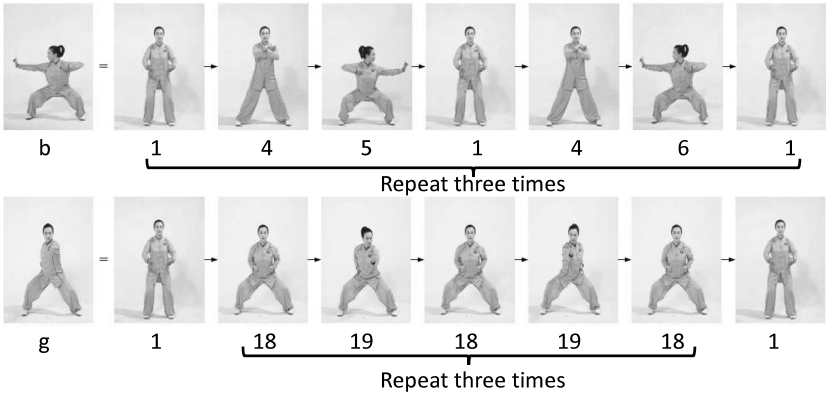


Figure 9. Example of decomposing BaDuanJin movements.

For a video of BaDuanJin movements, we first use a movement classifier to recognize the corresponding movement, and then use the pose detection model to detect the posture results for each frame, i.e., continuous posture labels. The sequence formed by these labels contains the period and speed of the movement, from which information on the synchronization and periodicity of different practitioners performing the same movement can be extracted.

During video processing, some frames may be missed or detected incorrectly, resulting in noise in the generated label sequence, which is not conducive to subsequent analysis and judgment. Therefore, it is necessary to conduct denoising. Since video frames are continuous, they do not suddenly change into another movement in the middle of a continuous movement. Based on this, noise can be filtered out. For example, if a few frames of movement 2 appear in a series of continuous movement labels 1, movement 2 is considered noise and will be removed.

Both the template movements and the test movements contain a certain regularity after processing and becoming label sequences. Either a template movement or a test movement is a periodic sequence. The difference lies in the length of the interval from one posture to the next one, which corresponds to different sequences and represents the difference in speed when different practitioners perform the same movement. To compare the similarity of label sequences, we propose a sequence pattern mining method based on sequence intervals.

The sequence pattern mining method is shown in Figure 10, where the template sequence and the test sequence are divided into multiple corresponding intervals. The associate pseudo-code of the algorithm is given in Algorithm 2. The elements in each interval are the same but they may have different lengths, representing the same movement state but with different durations. The formula for calculating the similarity between the test sequence and the template sequence is shown in Equation (10), where $\frac{|L(i)-L(i')|}{L(i')}$ represents the deviation of each corresponding interval.

$$Cycle\ Score = (1 - \frac{1}{n}\sum_{i=1}^{n}\frac{|L(i) - L(i')|}{L(i)}) \times 100 \tag{10}$$

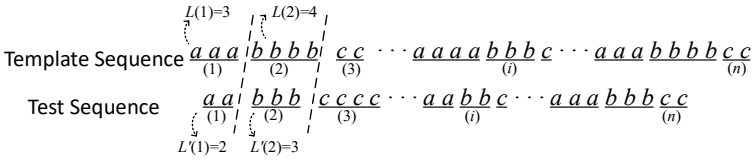


Figure 10. Process of calculating sequence periodic similarity.

Algorithm 2 Periodic similarity evaluation based on label sequence

Input:
Template label sequence: A
Test label sequence: B
Output:
Cycle Score
Divide template label sequence A into n intervals
Count the length of subsequence $L(1), L(2), ..., L(n)$
Divide test label sequence B into n intervals
Count the length of subsequence $L'(1), L'(2), ..., L'(n)$
$sum \leftarrow 0$
for $(i = 1; i < n; i++)$ **do**
$\theta_i \leftarrow |L(i) - L'(i)|/L(i)$
$sum \leftarrow sum + \theta_i$
end for
Cycle Score $\leftarrow (1 - sum/n) \times 100$

5.3. Decisions and Suggestions

We tested different exercisers, and associated suggestions were given to them. For example, one of the tests included two coaches and one beginner. By using Equation (10), it was calculated that the periodic similarity score between the beginner and the coach was 65.7. The periodic similarity score between coach one and coach two was 90.4. Since the overall speed of the beginner was too fast, it resulted in a lower periodic similarity score. The movement of both coach one and coach two had a similar speed, resulting in a higher periodic similarity score.

The scoring system was designed to have two perspectives: the similarity of movements and the periodicity of movements. Users can change their perspectives by adjusting a factor and obtaining suggestions accordingly. If they want to check whether their movements are performed correctly while periodicity is not particularly important, they can increase the weight of similarity. In this test, the weight factor is set to 0.4, which means the proportion of movement label similarity assessment counts 40%, and the proportion of joint vector similarity assessment counts the rest 60%.

The displacement trajectories of key points in the human body can also be visualized. Figure 11 depicts the changes in the right knee's x-coordinate of three testers. The trends of curves (a) and (c) are the same, indicating a higher similarity in movements. Curve (b) also has a similar pattern to curve (a), but the movement amplitude of curve (b) is very low, which results in a poor similarity of the curve. Those curves can be used to give suggestions to the exercisers.

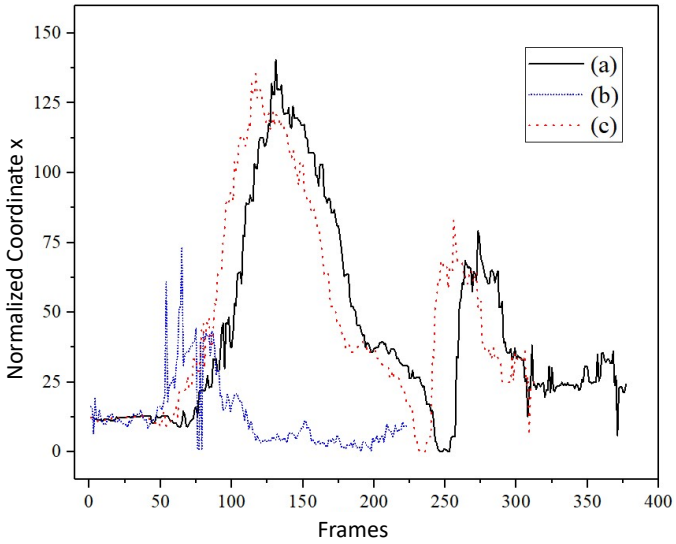


Figure 11. The changes of the right Knee's x-coordinate of three testers.

In order to present the system detection results more clearly and directly, we designed a visual interface using QT. This interface can help practitioners compare and correct their movements by displaying the predicted results of action techniques and analyzing joint angles, thereby assisting in their training. The movement recognition and scoring interface is shown in Figure 12.

Practitioners can upload a video of the BaDuanJin exercise by clicking the "Video Input" button. Then, by clicking the "Run" button, the system calls the trained action classification model to perform the detection and display the results in the window. Finally, on the right side, the category of the exercise video and the similarity score with the standard video are displayed. Based on this score, practitioners can have a preliminary understanding of their own level and enhance their interest in exercise.

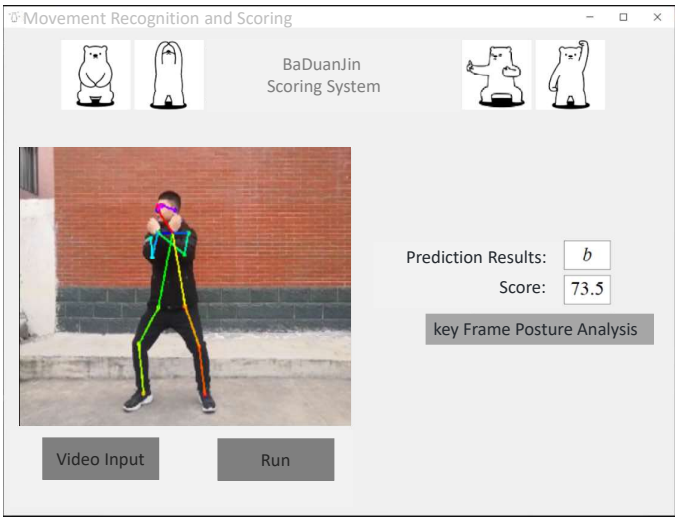
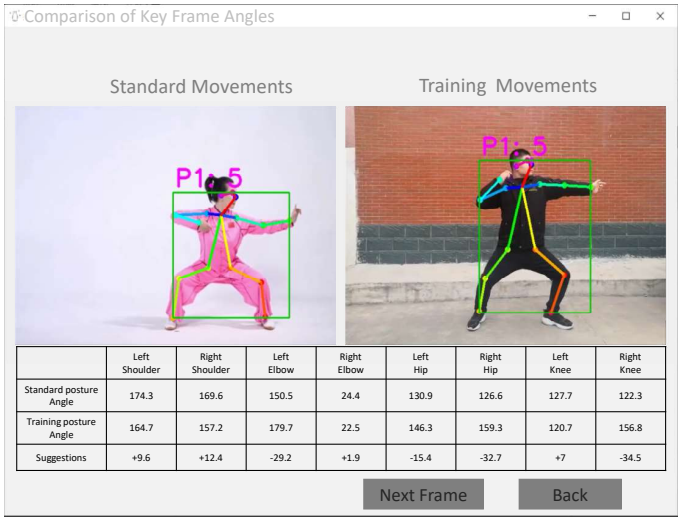


Figure 12. Movement recognition and scoring interface.

After obtaining the action results from the action classifier and the similarity score of the movements, practitioners can click on the key frame posture analysis results to view the key movement details of the corresponding technique. The evaluation interface is shown in Figure 13. The left window displays the posture frames of the standard action, and the right window displays the key frames of the input test video.



	Left Shoulder	Right Shoulder	Left Elbow	Right Elbow	Left Hip	Right Hip	Left Knee	Right Knee
Standard posture Angle	174.3	169.6	150.5	24.4	130.9	126.6	127.7	122.3
Training posture Angle	164.7	157.2	179.7	22.5	146.3	159.3	120.7	156.8
Suggestions	+9.6	+12.4	-29.2	+1.9	-15.4	-32.7	+7	-34.5

Figure 13. Joint angle comparison interface.

The scoring method for the BaDuanJin fitness movements designed in this paper measures the similarity of the practitioners' movements from two perspectives: overall movement similarity and the periodicity of the movements. This method achieves a quantitative assessment of the similarity when practitioners perform the BaDuanJin movements.

6. Conclusions

In this paper, we first introduce the lightweight backbone network Oct-MobileNet to overcome the limitations of traditional algorithms by employing octave convolution and attention mechanisms. This enables the extraction of high-frequency features from the human body contour, resulting in improved accuracy with reduced model computational burden. Furthermore, we introduce a novel approach for action recognition that combines skeleton-based information and multiple feature fusion. Additionally, we address the evaluation of traditional fitness exercises and propose a multimodal information-based

assessment method that combines pose detection and key point analysis. Overall, the methods presented in this paper offer promising solutions to the challenges of action recognition and assessment in the context of fitness training. The outcomes of the experiments demonstrate the effectiveness of the proposed approaches in accurately recognizing and evaluating human actions. By providing the feedback, fitness practitioners can improve their exercise performance and maximize the benefits derived from their fitness routines.

Author Contributions: Writing—original draft, L.W.; Writing—review & editing, B.S. and G.W.; Data analysis, Q.L., R.G. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

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

References

1. Lian, C.; Ma, R.; Wang, X.; Zhao, Y.; Peng, H.; Yang, T.; Zhang, M.; Zhang, W.; Sha, X.; Li, W.J. ANN-enhanced IoT wristband for recognition of player identity and shot types based on basketball shooting motion analysis. *IEEE Sens. J.* **2021**, *22*, 1404–1413. [CrossRef]
2. Albert, J.A.; Zhou, L.; Glöckner, P.; Trautmann, J.; Ihde, L.; Eilers, J.; Kamal, M.; Arnrich, B. Will You Be My Quarantine: A Computer Vision and Inertial Sensor Based Home Exercise System. In Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare, Atlanta, GA, USA, 18–20 May 2020; pp. 380–383.
3. Gupta, A.; Gupta, H.P. Yogahelp: Leveraging motion sensors for learning correct execution of yoga with feedback. *IEEE Trans. Artif. Intell.* **2021**, *2*, 362–371. [CrossRef]
4. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. *Sensors*, **2019**, *19*, 1005. [CrossRef] [PubMed]
5. Ji, R. Research on basketball shooting action based on image feature extraction and machine learning. *IEEE Access* **2020**, *8*, 138743–138751. [CrossRef]
6. Xu, J.; Song, R.; Wei, H.; Guo, J.; Zhou, Y.; Huang, X. A fast human action recognition network based on spatio-temporal features. *Neurocomputing*, **2021**, *441*, 350–358. [CrossRef]
7. Zou, L.; Pan, Z.; Yeung, A.; Talwar, S.; Wang, C.; Liu, Y.; Shu, Y.; Chen, X.; Thomas, G.A. A review study on the beneficial effects of Baduanjin. *J. Altern. Complement. Med.* **2018**, *24*, 324–335. [CrossRef] [PubMed]
8. Banjarey, K.; Sahu, S.P.; Dewangan, D.K. A survey on human activity recognition using sensors and deep learning methods. In Proceedings of the 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 8–10 April 2021; pp. 1610–1617.
9. Liu, Y.; Nie, L.; Liu, L.; Rosenblum, D.S. From action to activity: Sensor-based activity recognition. *Neurocomputing*, **2016**, *181*, 108–115. [CrossRef]
10. Pansiot, J.; Lo, B.; Yang, G.Z. Swimming stroke kinematic analysis with BSN. In Proceedings of the International Conference on Body Sensor Networks, Singapore, 7–9 June 2010; pp. 153–158.
11. Ma, H.; Liu, H. Research on human motion recognition system based on MEMS sensor network. In Proceedings of the IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 20–22 December 2019; Volume 1, pp. 2530–2534.
12. Shaeffer, D.K. MEMS inertial sensors: A tutorial overview. *IEEE Commun. Mag.* **2013**, *51*, 100–109. [CrossRef]
13. Safaric, S.; Malaric, K. ZigBee wireless standard. In Proceedings of the IEEE International Symposium on Electronics in Marine, Zadar, Croatia, 7–10 June, 2006; pp. 259–262.
14. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [CrossRef]
15. Choudhury, N.A.; Moulik, S.; Choudhury, S. Cloud-based real-time and remote human activity recognition system using wearable sensors. In Proceedings of the IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 28–30 September 2020; pp. 1–2.
16. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
17. Shao, Z.; Guo, J.; Zhang, Y.; Zhu, R.; Wang, L. LightBGM for human activity recognition using wearable sensors. In Proceedings of the International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Xi'an, China, 27–28 March 2021; pp. 668–671.
18. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of Advances in Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

19. Sun, G. Golf Swing Correction Based on Deep Learning Body Posture Recognition. In Proceedings of the International Conference on Pattern Recognition and Intelligent Systems, Bangkok, Thailand, 28–30 July 2021; pp. 72–76.
20. Xu, C.; Fu, Y.; Zhang, B.; Chen, Z.; Jiang, Y.G.; Xue, X. Learning to score figure skating sport videos. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4578–4590. [CrossRef]
21. Yang, R.; Tao, R.; Wang, Z.; Feng, X. Etiquette Action Similarity Evaluation Based on Posture Recognition. In *Knowledge Management in Organizations, In Proceedings of the 15th International Conference, Kaohsiung, Taiwan, 20–22 July 2021*; Springer: Cham, Switzerland, 2021; pp. 404–415.
22. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2334–2343.
23. 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, UT, USA, 18–22 June 2018; pp. 7103–7112.
24. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef] [PubMed]
25. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Bottom-up higher-resolution networks for multi-person pose estimation. *arXiv* **2019**, arXiv:1908.10357.
26. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]