*future internet*

MDPI

*Article*

# Comparative Analysis of Skeleton-Based Human Pose Estimation

**Jen-Li Chung, Lee-Yeng Ong * and Meng-Chew Leow**

Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, Melaka 75450, Malaysia
* Correspondence: lyong@mmu.edu.my

**Abstract:** Human pose estimation (HPE) has become a prevalent research topic in computer vision. The technology can be applied in many areas, such as video surveillance, medical assistance, and sport motion analysis. Due to higher demand for HPE, many HPE libraries have been developed in the last 20 years. In the last 5 years, more and more skeleton-based HPE algorithms have been developed and packaged into libraries to provide ease of use for researchers. Hence, the performance of these libraries is important when researchers intend to integrate them into real-world applications for video surveillance, medical assistance, and sport motion analysis. However, a comprehensive performance comparison of these libraries has yet to be conducted. Therefore, this paper aims to investigate the strengths and weaknesses of four popular state-of-the-art skeleton-based HPE libraries for human pose detection, including OpenPose, PoseNet, MoveNet, and MediaPipe Pose. A comparative analysis of these libraries based on images and videos is presented in this paper. The percentage of detected joints (PDJ) was used as the evaluation metric in all comparative experiments to reveal the performance of the HPE libraries. MoveNet showed the best performance for detecting different human poses in static images and videos.

**Keywords:** human pose estimation; OpenPose; PoseNet; MoveNet; MediaPipe Pose

**Citation:** Chung, J.-L.; Ong, L.-Y.; Leow, M.-C. Comparative Analysis of Skeleton-Based Human Pose Estimation. *Future Internet* **2022**, *14*, 380. https://doi.org/10.3390/fi14120380

Academic Editor: Manuel José Cabral dos Santos Reis

Received: 28 October 2022
Accepted: 9 December 2022
Published: 15 December 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



## 1. Introduction

Human pose estimation (HPE) aims to locate all of the human body parts from input images or videos. Nowadays, HPE has become a popular task in the field of computer vision. It is widely used in video surveillance [1–6], medical assistance [7–15], and sport motion analysis [16–28]. Human keypoints are used to classify the poses and measure the correctness of poses in these applications. Using an intelligent video surveillance system, the human keypoints can be extracted from human body parts to classify poses between kidnapping and child abuse cases. In the aspect of medical assistance, detected keypoints from body parts can be used to evaluate the correctness of postures for physiotherapy exercises, fall detection, and in-home rehabilitation. In addition, the performance and correctness of an athlete's movements can be evaluated by comparing the detected keypoints from body parts with reference poses (ground truth).

HPE can be classified into two-dimensional (2D) HPE and three-dimensional (3D) HPE. It can also be classified into single-person HPE and multi-person HPE based on the number of people captured in the input image. In both single-person and multi-person HPE, it can be further classified into top-down and bottom-up methods, based on the ways of detecting the skeleton keypoints [29]. This paper focuses on a comparative study of 2D single-person HPE.

As the demand for HPE increases, many skeleton-based HPE algorithms have been developed and packaged into libraries to provide ease of use for researchers. The performance of these HPE libraries is important to ensure the reliability of the different practical

applications for which they are integrated. For instance, when the HPE library is applied to an in-home rehabilitation system, it needs to accurately detect the poses of patients freely performing rehabilitation poses in different home environments to ensure the reliability of the application. This situation is even more complicated when common challenges, such as inappropriate camera position and self-occlusion [11,20,22,23], affect skeleton keypoint detection. Recently, four state-of-the-art HPE libraries have been applied in various applications, namely PoseNet [30], MoveNet [31], OpenPose [32], and MediaPipe Pose [33]. Table 1 shows a list of applications in different domains that have utilized these four HPE libraries in the last 5 years.

**Table 1.** Applications of PoseNet, MoveNet, OpenPose, and MediaPipe Pose in different domains.

| Domain | HPE Library | Year | Purpose of Application |
|---|---|---|---|
| Video Surveillance | OpenPose [6] | 2018 | Kidnapping detection–using HPE to classify kidnapping cases and normal cases in an intelligent video surveillance system. |
| | OpenPose [5] | 2019 | A child abuse prevention decision-support system–using OpenPose to classify adults and children in CCTV. |
| Medical Assistance | OpenPose [15] | 2020 | A fall detection system–using OpenPose to extract features of human body. |
| | PoseNet [8] | 2020 | Automatic feedback on incorrect posture for physiotherapy exercises. |
| | PoseNet [10] | 2021 | A telehealth system providing in-home rehabilitation. |
| | OpenPose [11] | 2021 | Measure joint angles and conduct semi-automatic ergonomic postural assessments to evaluate the risk of musculoskeletal disorders. |
| | MoveNet [14] | 2021 | A healthcare system that measures patient's strength, balance, and range of motion during physical therapy activities. |
| | MediaPipe Pose [12] | 2022 | A fall detection system. |
| | MediaPipe Pose [13] | 2022 | A posture corrector system–to notify people who are spending most of their time sitting in front of the computer with bad posture to avoid long-term health issues. |
| Sport Motion Analysis | OpenPose [28] | 2018 | A basketball free-throw shooting prediction system–using OpenPose to generate body keypoints. |
| | OpenPose [21] | 2020 | A real-time push-up counter–to classify the correct and incorrect push-ups. |
| | OpenPose [20] | 2021 | A system to evaluate baseball swinging poses and help baseball players correct their poses. |
| | MediaPipe Pose [24] | 2021 | A mobile application–to analyze, improve, and track cricket players' batting performance. |
| | PoseNet [25] | 2021 | A real-time workout analyzer–allows fitness enthusiast to perform their workout accurately at home and with proper guidance. |
| | PoseNet [26] | 2021 | A fitness tutor–to maintain the correctness of the posture during workout exercises. |
| | PoseNet [27] | 2021 | A fitness application–provides instant feedback to users to ensure the accuracy of their workout exercise poses. |
| | MediaPipe Pose [22] | 2022 | To score the human body's balance ability on the wobble board. |
| | MediaPipe Pose [23] | 2022 | A free weight exercise tracking software–allows users to learn and correct their exercise poses. |

The four HPE libraries face two common challenges in pose estimation, including inappropriate camera position and the self-occlusion effect. Even though each HPE library uses different approaches to overcome these challenges, the strengths and weaknesses of these four existing HPE libraries have yet to be discovered. Therefore, the comparative performance of these libraries should be carried out to investigate their robustness in detecting different human poses. This paper aims to compare the performance of these four

state-of-the-art HPE libraries for human pose detection and analyze the strengths and weaknesses of each HPE library. Hence, a comparative analysis of these four HPE libraries based on images and videos was carried out. To the best of the authors' knowledge, this paper is the first attempt to compare and analyze the performance of PoseNet, OpenPose, MoveNet, and MediaPipe Pose using both image and video datasets.

The rest of this paper is organized as follows. Section 2 reviews the existing comparative analysis of the different HPE libraries and summarizes the functionalities of PoseNet, OpenPose, MoveNet, and MediaPipe Pose. The methodology used to evaluate the performance of each HPE library is presented in Section 3. Next, Section 4 provides the experimental results, in terms of image and video datasets. Section 5 analyzes the strengths and weaknesses of each HPE library. The last section presents the conclusion of the study.

## 2. Literature Review

### *2.1. Existing Comparative Analysis*

In two of the existing comparative analysis studies, the researchers compared the performance of a few HPE libraries using image datasets. In [34], the AR dataset was used to compare the performance of OpenPose and BlazePose [33] using PCK@0.2 as the evaluation metric. The results showed that OpenPose achieved slightly better performance than BlazePose, with the results of 87.8 and 84.1, respectively. Ref. [31] used OpenPose, PoseNet, MoveNet Lightning, and MoveNet Thunder in their study. The researchers used two image datasets: the COCO [35] and MPII [36] datasets. The performances of the HPE libraries were measured using their own proposed evaluation metric. PoseNet had the best performance, while MoveNet Lightning had the worst performance. The comparative analysis of both existing studies was limited to a few selections of HPE libraries using image datasets. However, the performance of four state-of-the-art HPE libraries based on video datasets has yet to be investigated. In this paper, the authors have conducted the experiments using image and video datasets.

### *2.2. HPE Libraries*

Four state-of-the-art HPE libraries are discussed in this section and their specifications are summarized in Table 2. Among the four HPE libraries, the total number of commonly detected keypoints is 17. The commonly detected keypoints of the head include ears, eyes, and nose (5 keypoints). The 6 commonly detected keypoints of the shoulders, elbows, and wrists are categorized as the upper body, while the lower body includes 6 keypoints from the hips, knees, and ankles. In addition, OpenPose and MediaPipe Pose provide more annotations of the keypoints at the face, hand, and foot to reach the maximum number of keypoints, at 135 and 33 keypoints, respectively. OpenPose provides an additional 70 keypoints of the face, 20 keypoints of both hands, 1 keypoint of the upper body, and 7 keypoints of the lower body. MediaPipe Pose provides 6 additional keypoints of the head, 6 keypoints of the upper body, and 4 keypoints of the lower body. The approach of keypoint detection in the HPE libraries can be classified into top-down and bottom-up methods. In the top-down method, the number of people is first detected from the given input and each person is assigned into a separate bounding box, respectively [37]. Subsequently, the keypoint estimation is performed in each bounding box. In contrast to the top-down method, the bottom-up method performs keypoint detection in the first step [38]. After that, the keypoints are grouped based on human instances. Among these four libraries, PoseNet and MediaPipe Pose employ the top-down method while OpenPose and MoveNet use the bottom-up method to perform human pose estimations. The four HPE libraries use different underlying networks for pose estimation. OpenPose uses ImageNet with the VGG-19 backbone, PoseNet uses ResNet [39] and MobileNet [40], MediaPipe Pose uses the Convolutional Neural Network (CNN), and MoveNet uses the MobileNetV2.

**Table 2.** Specifications of each HPE library.

| HPE Libraries | Released Year | Maximum Number of Keypoints | Keypoints Position in the Body Parts | Type of Pose | Method | Underlying Network |
|---|---|---|---|---|---|---|
| OpenPose [32] | 2017 | 135 | Face, hand, head, upper body, lower body | Single- and multi-person | Bottom-up | ImageNet with VGG-19 |
| PoseNet [30] | 2017 | 17 | Head, upper body, lower body | Single- and multi-person | Top-down | ResNet and MobileNet |
| MediaPipe Pose [33] | 2020 | 33 | Head, upper body, lower body | Single-person | Top-down | CNN |
| MoveNet [31] | 2021 | 17 | Head, upper body, lower body | Single- and multi-person | Bottom-up | MobileNetV2 |

OpenPose is the first open-source library available since 2017 for 2D multi-person HPE [32]. OpenPose employs a non-parametric representation known as Part Affinity Fields (PAFs) to detect the body parts associated with the person in an input image. The PAFs describe a list of 2D vector fields in an image, encoding both orientation and location of the body limits. In 2019, Cao et al. [41] released a new version of OpenPose that combined body and foot keypoint detectors. The combined detector needs less inference time than running the body and foot keypoint detectors independently, while also maintaining the accuracy rate. Hence, OpenPose became the first open-source library that can detect body, hand, foot, and facial keypoints on a single image with a total of 135 keypoints. Furthermore, OpenPose is also able to perform the task of vehicle keypoint detection by utilizing the same network architecture [41].

Similar to OpenPose, PoseNet was also released in 2017 [30]. It was built on a TensorFlow machine learning platform and provides real-time HPE implementation in the browser. There are two versions of the algorithm in PoseNet. One algorithm is used to estimate the single pose and the other is used to estimate multiple poses from the input image or video. Both algorithms are able to detect 17 keypoints in a single person. The computational time of the multi-person HPE algorithm is slightly slower than that of the single-person HPE algorithm. However, it is not affected by the number of detected persons. When using the single-person algorithm, the keypoints might be conflated if there is more than one person in the input image or video. Moreover, there are two architectures in PoseNet, which are ResNet [39] and MobileNet [40]. MobileNet is designed for mobile devices. It is more lightweight than the ResNet, but has lower accuracy. Although ResNet achieves higher accuracy than MobileNet, its larger number of layers requires longer loading and inference time.

In 2020, a solution called MediaPipe Pose was released to achieve higher fidelity human body pose tracking using the machine learning approach [33]. It utilizes the BlazePose and ML Kit Pose Detection API to infer a maximum of 33 keypoints (3D landmarks) from an RGB input. It can be performed in real-time on mobile phones, desktops, or laptops. BlazePose employs a two-step detector-tracker pipeline for single-person pose estimation [33]. The first step of this pipeline locates the region-of-interest (ROI) of the person inside the image frame. Subsequently, the tracker uses the ROI from the detector as the input to predict the position of each keypoint within the ROI. If the input is a video, the detector is invoked at the first frame to extract the human ROI, followed by keypoint extraction using the tracker. The tracker uses the same ROI to estimate the keypoints of the human in the next frame. When the algorithm loses track of the person, the detector is invoked again to generate a new ROI.

MoveNet [31], which was released in 2021, is a pose detection model that detects 17 keypoints of a single person in real-time. There are two variants of MoveNet, which are Lightning and Thunder. The accuracy of Lightning is lower than that of Thunder. However, the inference time of Lightning is faster than that of Thunder. MoveNet uses heatmaps to accurately localize the human keypoints. Its architecture consists of two components, which are a feature extractor and a set of prediction heads. The prediction

technique of MoveNet loosely follows that of CenterNet [42] to improve its accuracy and speed. CenterNet is an object detector that uses the keypoint estimation networks to find the center points and regress to the object size, location, and orientation. The feature extractor is MobileNetV2 with an attached feature pyramid network [43] to produce a high resolution and semantically rich feature map output. MobileNetV2 is a neural network designed with mobile devices to extract features for object detection, classification, and semantic segmentation. There are four parts in the prediction heads, which include person center heatmap, keypoint regression field, person keypoint heatmap, and 2D per-keypoint offset field. They are responsible for predicting the human keypoints using heatmaps.

## 3. Methodology

In this section, the descriptions of the datasets used in the experiments are presented. After selecting the dataset, data pre-processing was carried out. The HPE procedure was then performed. Next, an evaluation metric was used to evaluate the performance of the HPE library. The evaluation metric is discussed in this section.

### *3.1. Datasets*

The datasets of image and video sources used in this experiment were the Microsoft Common Object in Context (COCO) [35] and Penn Action [44] datasets, respectively. Table 3 shows the characteristics of the COCO and Penn Action datasets. Both datasets have 6 common upper body keypoints and 6 lower body keypoints. However, the difference between these two datasets is the number of keypoints annotated for the head. COCO has 5 keypoints, including the nose, eyes, and ears, while Penn Action only provides 1 keypoint at the head position. Figures 1 and 2 show sample images of the COCO and Penn Action datasets, respectively.

COCO [35] is a large-scale object detection, segmentation, and captioning dataset. It is commonly used in experiments for HPE [45–47]. It consists of 330,000 images, 1.5 million object instances, 80 object categories, 91 stuff categories, and 250,000 humans with keypoints. This dataset provides annotations for the body keypoint detection, where each instance of a person is labeled with 17 keypoints. There are various versions of COCO. COCO 2017 is commonly selected for HPE experiments.

Penn Action [44] is a video dataset, which consists of 2326 video sequences with 15 actions. It is commonly used in HPE experiments [18,48–50]. Each video sequence contains RGB image frames and annotations. The annotations include different human actions, 2D bounding boxes to locate human positions, and skeleton keypoints in the body parts. Each instance of a human is labeled with 13 keypoints.

**Table 3.** The characteristics of the COCO and Penn Action datasets.

| Dataset Name | Dataset Type | Number of Keypoints Per Person | Annotation of Body Parts Provided |
|---|---|---|---|
| COCO | Image | 17 | Nose, Left Eye, Right Eye, Left Ear, Right Ear, Left Shoulder, Right Shoulder, Left Elbow, Right Elbow, Left Wrist, Right Wrist, Left Hip, Right Hip, Left Knee, Right Knee, Left Ankle, Right Ankle |
| Penn Action | Video | 13 | Head, Left Shoulder, Right Shoulder, Left Elbow, Right Elbow, Left Wrist, Right Wrist, Left Hip, Right Hip, Left Knee, Right Knee, Left Ankle, Right Ankle |

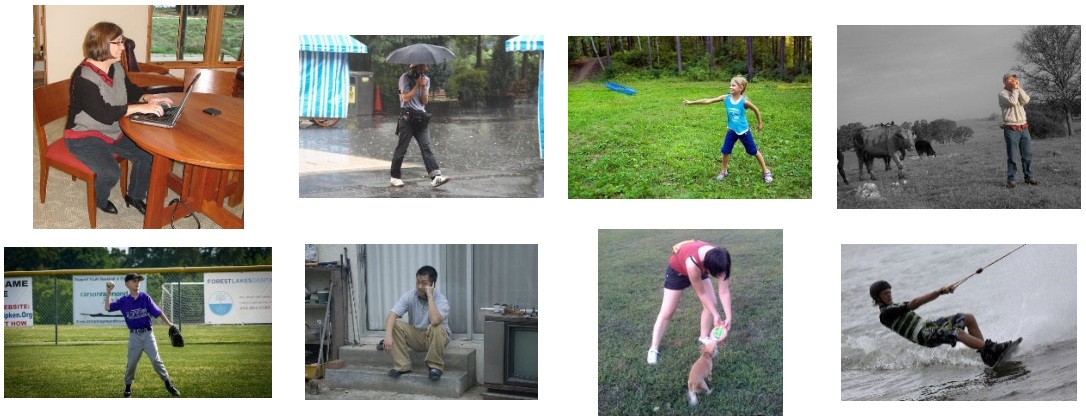
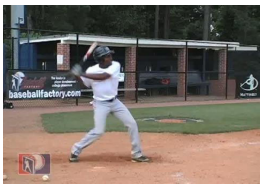


**Figure 1.** Sample images from the COCO dataset.

(a) Action 1: Baseball pitch


(b) Action 2: Bench press




(c) Action 3: Baseball swing

(d) Action 4: Bowling

(e) Action 5: Clean and jerk

(f) Action 6: Golf swing

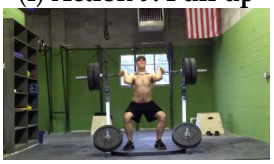
(g) Action 7: Jump rope

(h) Action 8: Jumping jacks

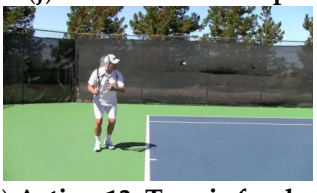
(i) Action 9: Pull up

(j) Action 10: Push up

(k) Action 11: Sit up



(l) Action 12: Squat



(m) Action 13: Tennis forehand

(n) Action 14: Tennis serve

(o) Action 15: Strum guitar

**Figure 2.** Sample video frames of each action from the Penn Action dataset.

### *3.2. Data Pre-Processing*

Before evaluating the performance of the HPE libraries, data pre-processing was conducted to filter out irrelevant data in both datasets. There are three types of images in COCO, including images of a single person, images of multiple-people, and images without people. This experiment focused on single-person HPE. Hence, images with multiple

people and without people were removed. In addition, images with only half of the human body were removed. Thus, 1100 remaining images were used in the experiment. In order to compare the performance of the four HPE libraries, 17 commonly detected keypoints from the human body were matched with 17 annotations provided by the dataset (ground truth).

For the Penn Action videos, the action of guitar strumming was removed since the video frames only consisted of the upper half of the human body. The first 14 actions were used in this experiment (refer to Figure 2). Since the Penn Action dataset only provided 1 annotation of a keypoint of the head, which differed from the four HPE libraries (refer to Table 3), the head annotation was removed from the experiments to maintain a fair comparison among all libraries. Thus, the 12 remaining keypoints were used as the ground truth.

### *3.3. Evaluation Metrics*

Evaluation metrics play an important role in evaluating the quality of HPE libraries. The evaluation metric used in this study was the percentage of detected joints (PDJ), which was able to measure the performance of the HPE library [37,51,52]. PDJ uses the Euclidean distance between the ground truth and predicted keypoints in pixel(s) to measure the detection accuracy of the HPE libraries. The higher the value of PDJ, the higher the accuracy rate. The calculation of Euclidean distance $d(x, y)$ between the ground truth $(x_1, y_1)$ and predicted keypoints $(x_2, y_2)$ is shown in Equation (1). The threshold of PDJ was 0.05 for the value of the torso diameter. The torso diameter was computed for the Euclidean distance from the left shoulder to the right hip, represented as the coordinates $(x_{ls}, y_{ls})$ and $(x_{rh}, y_{rh})$, as shown in Equation (2). When the $d(x, y)$ between the predicted keypoints and ground truth keypoints was smaller than the threshold, the predicted keypoints were considered to be correctly detected. Hence, the PDJ can be deduced as shown in Equation (3), where $n$ represents the total number of predicted joints.

$$d(x, y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \text{ (pixel)} \quad (1)$$

$$torso\ diameter = \sqrt{(x_{ls} - x_{rh})^2 + (y_{ls} - y_{rh})^2} \quad (2)$$

$$PDJ = \frac{\sum_{i=1}^{n} bool(d_i < 0.05 \times torso\ diameter)}{n} \quad (3)$$

## 4. Experiment Results

Three experiments were conducted to evaluate the performances of the HPE libraries. For the image dataset, an experiment was conducted to compare the performance of each HPE library for each image. For the video dataset, the first experiment was conducted to compare the performance of each HPE library for each video frame, whereas the second experiment investigated how well each HPE library performed for each body part of each action.

### *4.1. Image Dataset*

The PDJ value for each image was calculated to compare the performance of the HPE libraries for each image. The results are presented in a box plot, as shown in Figure 3. Among the four HPE libraries, MoveNet (orange box) achieved the highest PDJ in terms of the lower fence, first quartile, median, and third quartile values. The second best performing HPE library was OpenPose (blue box), which achieved the same median and third quartile values as MoveNet; however, OpenPose had lower first quartile and fence (outlier) values. The third best performing HPE library was PoseNet (green box), followed by MediaPipe Pose (red box). The minimum values of MediaPipe Pose and PoseNet were 0. Meanwhile, the outlier values for MoveNet and OpenPose were also 0. The value of 0

indicated that the keypoints detected in some of the images were incorrectly matched with the ground truth provided by the COCO dataset.

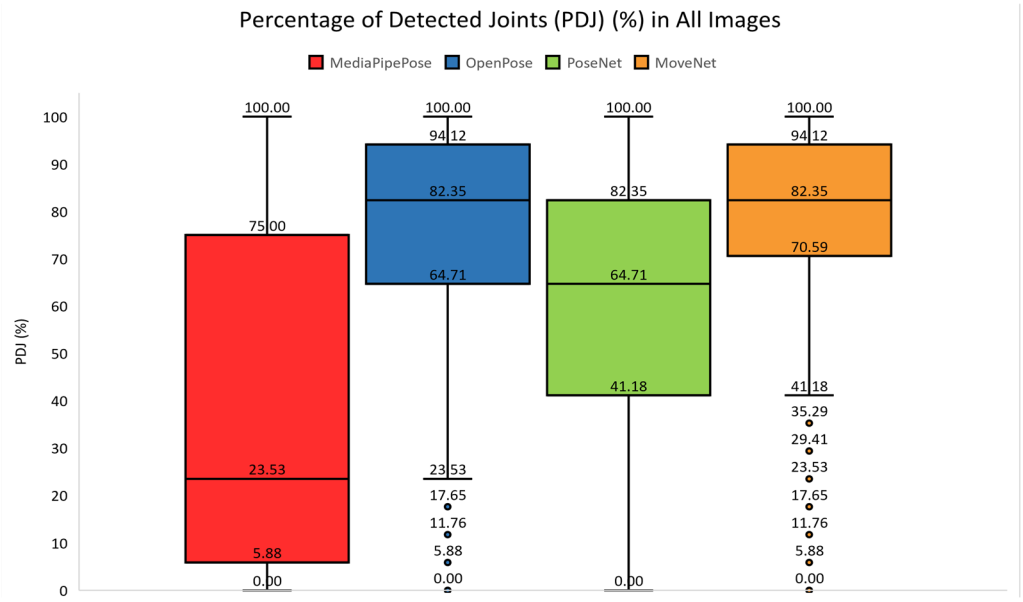



**Figure 3.** Box plot of PDJ of all images in the image dataset.

Table 4 divides the performance of the HPE libraries into five groups, including 0%, 0–25%, 25–50%, 50–75%, and 75–100%. Each group shows the number of images that were recognized by the HPE library in the specific range of PDJ values. MoveNet had the least number of images that achieved 0% of PDJ, which included 5 images. OpenPose achieved the second lowest number of images at 0% (8 images), followed by PoseNet (30 images) and MediaPipe Pose (240 images). In the range of 0–25%, MoveNet had the least number of images (33 images), followed by OpenPose (87 images), PoseNet (128 images), and MediaPipe Pose (342 images). Likewise, MoveNet had the least number of images that achieved the PDJ within the range of 25–50% (68 images).

MoveNet achieved superior performance because it had the highest number of images in the last two groups (50–100%), which indicated that more than 50% of the detected keypoints from 994 images were correctly matched with the ground truth. In contrast, MediaPipe Pose was found to have the poorest performance because it had the highest number of images in the first to third groups (0–50%). In a total of 698 images, less than 50% of the detected keypoints were correctly matched with the ground truth. Overall, MoveNet achieved the top performance because it showed the highest number of images in the fifth group, which indicated that 747 out of 1,100 images were recognized by MoveNet with 75–100% detected keypoints. In contrast, MediaPipe Pose showed the poorest performance as it only achieved the lowest number of images in the range of 75–100%, but it received the highest number of images in the 0% group. OpenPose achieved the second highest performance, which was slightly lower than MoveNet. PoseNet and MediaPipe Pose achieved the third and fourth highest performances, respectively.

**Table 4.** Number of images recognized by each HPE library in each specific range of PDJ values.

| | **0%** | **0% < PDJ ≤ 25%** | **25% < PDJ ≤ 50%** | **50% < PDJ ≤ 75%** | **75% < PDJ ≤ 100%** |
|---|---|---|---|---|---|
| MediaPipe Pose | 240 | 342 | 116 | 127 | 275 |
| OpenPose | 8 | 87 | 85 | 252 | 668 |
| PoseNet | 30 | 128 | 185 | 323 | 434 |
| MoveNet | 5 | 33 | 68 | 247 | 747 |

In the overall comparison, MoveNet was the most robust because it achieved the top performance in terms of lower fence, first quartile, median, and third quartile values compared to the other HPE libraries. MoveNet achieved more than 50% PDJ value for 994 out of 1100 images. In addition, MediaPipe Pose showed the poorest performance in the image dataset as it has the lowest PDJ in terms of first quartile, median, and third quartile values. MediaPipe Pose also showed that less than 50% of detected keypoints could be correctly matched with the ground truth in 698 out of 1100 images. OpenPose and PoseNet achieved the second best and third best performances, respectively.

### *4.2. Video Dataset*

The mean PDJ for each action was calculated and the PDJ values for all actions are shown in Figure 4. Among three HPE libraries, MoveNet (orange box), MediaPipe Pose (red box), and PoseNet (green box) achieved slightly similar performances in terms of minimum, first quartile, median, maximum, and third quartile values. MediaPipe Pose (red box) had the highest PDJ in terms of maximum and median values, while MoveNet (orange box) scored the highest in terms of minimum, first, and third quartile values. OpenPose (blue box) portrayed a weak performance with the lowest values for all quartiles.

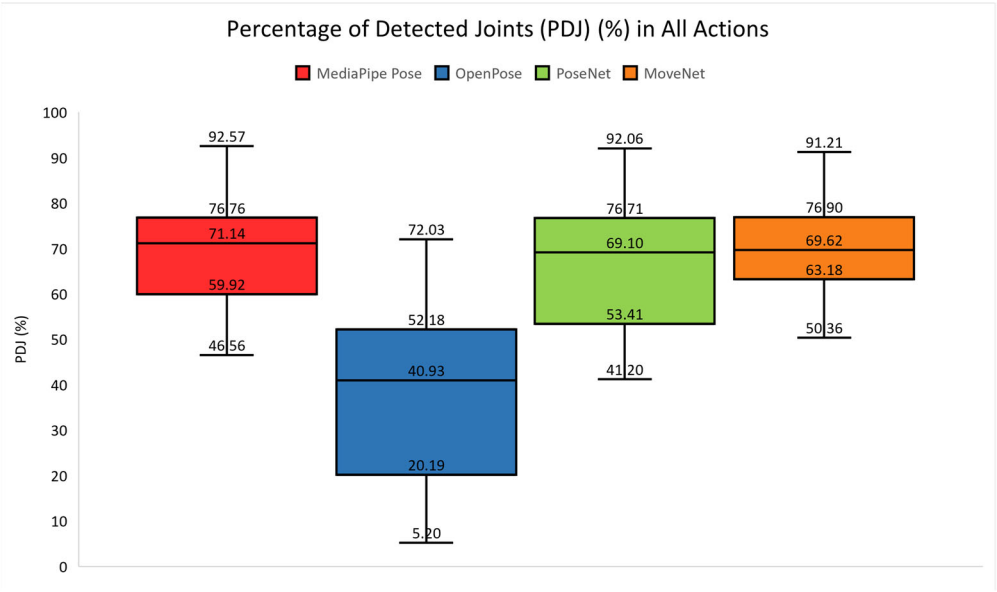



**Figure 4.** PDJ in all actions in the video dataset.

Similar to Table 4, Table 5 lists the performance of the HPE libraries for all actions into five groups, which are 0%, 0–25%, 25–50%, 50–75%, and 75–100%. There was no action showing 0% in all HPE libraries. In the range of 75–100%, MediaPipe Pose achieved the greatest number of actions, which was 5 actions. MoveNet and PoseNet achieved 4 actions in this group, while OpenPose showed 0 actions in this group. In the fourth group (50–75%), MoveNet achieved the greatest number of actions (10 actions), followed by PoseNet (7 actions), MediaPipe Pose (6 actions), and OpenPose (4 actions). In the first three groups (0–50%), OpenPose showed the greatest number of actions (10 actions), MediaPipe Pose and MoveNet showed 3 actions, while MoveNet showed 0 actions in these groups. Overall, MoveNet achieved the best performance because it achieved above 50% PDJ in all actions. OpenPose showed the worst performance since the PDJ for 10 actions were lower than 50%.

**Table 5.** Number of actions recognized by each HPE library in each specific range of PDJ values.

| | 0% | 0% < PDJ ≤ 25% | 25% < PDJ ≤ 50% | 50% < PDJ ≤ 75% | 75% < PDJ ≤ 100% |
|---|---|---|---|---|---|
| MediaPipe Pose | 0 | 0 | 3 | 6 | 5 |
| OpenPose | 0 | 5 | 5 | 4 | 0 |
| PoseNet | 0 | 0 | 3 | 7 | 4 |
| MoveNet | 0 | 0 | 0 | 10 | 4 |

To get a better understanding of the performance of the HPE libraries for each action, the highest and lowest average PDJ values of each library for each action are highlighted in Table 6. Among 14 actions, the libraries achieved the best performance in Action 8 (jumping jacks). MediaPipe Pose showed the poorest performance in Action 12 (squat) among all actions. Coincidentally, both OpenPose and PoseNet showed the poorest performance in Action 2 (bench press). On the other hand, MoveNet showed the poorest performance in Action 4 (bowling). The overall average PDJ values for all actions are highlighted in the last row of Table 6. The performance rank from the highest to lowest PDJ values was MoveNet, MediaPipe Pose, PoseNet, and OpenPose. The performances among MoveNet, MediaPipe Pose, and PoseNet fell between 65% and 70%. The performance of OpenPose was much lower than the others, which fell at approximately 37%. Although MediaPipe Pose successfully detected 5 actions (refer Table 5), achieving between 75% and 100% PDJ, which was more than MoveNet, its overall performance was 1.38% lower than that of MoveNet.

**Table 6.** Overall average of PDJ (%) in each action.

| | MediaPipe Pose | OpenPose | PoseNet | MoveNet |
|---|---|---|---|---|
| Action 1 | 66.62 | 39.94 | 68.47 | 68.70 |
| Action 2 | 48.38 | 5.20 | 41.20 | 58.68 |
| Action 3 | 74.21 | 52.35 | 77.55 | 80.58 |
| Action 4 | 48.92 | 21.04 | 56.69 | 53.95 |
| Action 5 | 76.34 | 45.44 | 72.34 | 75.67 |
| Action 6 | 79.00 | 54.71 | 77.49 | 81.42 |
| Action 7 | 72.10 | 44.17 | 69.72 | 69.31 |
| Action 8 | 92.57 | 72.03 | 92.06 | 91.21 |
| Action 9 | 78.03 | 35.98 | 76.45 | 71.85 |
| Action 10 | 63.59 | 14.11 | 48.66 | 66.73 |
| Action 11 | 66.32 | 17.63 | 54.89 | 64.68 |
| Action 12 | 46.56 | 23.93 | 48.97 | 50.36 |
| Action 13 | 75.72 | 52.12 | 72.99 | 74.79 |
| Action 14 | 70.17 | 41.91 | 65.82 | 69.93 |
| Average PDJ (%) | 68.47 | 37.18 | 65.95 | 69.85 |

Green words represent the highest PDJ of each HPE library among all actions. Red words represent the lowest PDJ of each HPE library among all actions. The green cell represents the highest value in average PDJ. The red cell represents the lowest values in average PDJ.

The ranking of each library for each action is listed in Table 7. MediaPipe Pose achieved the top performance in 7 of 14 actions. It achieved the second highest performance in 3 actions and the third highest performance in 4 actions. MoveNet achieved the highest performance in 6 actions, second highest performance in 5 actions, and third highest performance in 3 actions. PoseNet showed the highest performance in only Action 4 (bowling), second highest performance in 6 actions, and third highest performance in 7 actions. OpenPose showed the lowest performance in all actions among all the HPE libraries.

**Table 7.** Overall average of PDJ (%) in each action.

| Action | Rank 1 | Rank 2 | Rank 2 | Rank 3 |
|---|---|---|---|---|
| Action 1 | MoveNet | PoseNet | MediaPipe Pose | OpenPose |
| Action 2 | MoveNet | MediaPipe Pose | PoseNet | OpenPose |
| Action 3 | MoveNet | PoseNet | MediaPipe Pose | OpenPose |
| Action 4 | PoseNet | MoveNet | MediaPipe Pose | OpenPose |
| Action 5 | MediaPipe Pose | MoveNet | PoseNet | OpenPose |
| Action 6 | MoveNet | MediaPipe Pose | PoseNet | OpenPose |
| Action 7 | MediaPipe Pose | PoseNet | MoveNet | OpenPose |
| Action 8 | MediaPipe Pose | PoseNet | MoveNet | OpenPose |
| Action 9 | MediaPipe Pose | PoseNet | MoveNet | OpenPose |
| Action 10 | MoveNet | MediaPipe Pose | PoseNet | OpenPose |
| Action 11 | MediaPipe Pose | MoveNet | PoseNet | OpenPose |
| Action 12 | MoveNet | PoseNet | MediaPipe Pose | OpenPose |
| Action 13 | MediaPipe Pose | MoveNet | PoseNet | OpenPose |
| Action 14 | MediaPipe Pose | MoveNet | PoseNet | OpenPose |

Red cells represent MediaPipe Pose. Orange cells represent MoveNet. Green cells represent PoseNet. Blue cells represent OpenPose.

Since the results in Table 6 show that the four HPE libraries achieved the best performance in Action 8 (jumping jacks) among all actions, a closer analysis was performed. MediaPipe Pose, MoveNet, and PoseNet reached approximately 92% while OpenPose achieved approximately 72% for Action 8. Figure 5 shows the video frame and detection results of Action 8, where the two challenges, self-occlusion and inappropriate camera position, are absent. Hence, the performance of HPE should be better when there are less challenges affecting keypoint detection. On the other hand, the performance of HPE is reduced when there are more challenges. MediaPipe Pose showed the poorest performance in Action 12 (squat); OpenPose and PoseNet showed the worst performance in Action 2 (bench press); while MoveNet showed the poorest performance in Action 4 (bowling). Figures 6–8 show the sample video frames and detection results of each HPE library for Actions 12, 2, and 4, respectively.

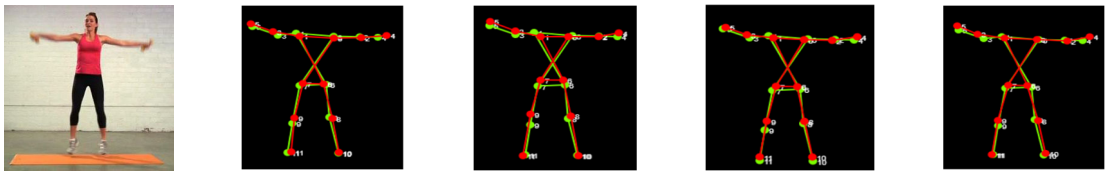



**Figure 5.** Sample video frame and detection result of Action 8 (jumping jacks). Green lines indicate the ground truth, red lines indicate the tested HPE library.

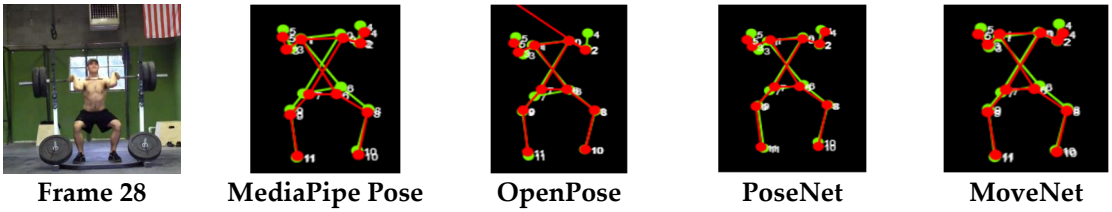



**Figure 6.** Sample video frame and detection result of Action 12 (squat). Green lines indicate the ground truth, red lines indicate the tested HPE library.

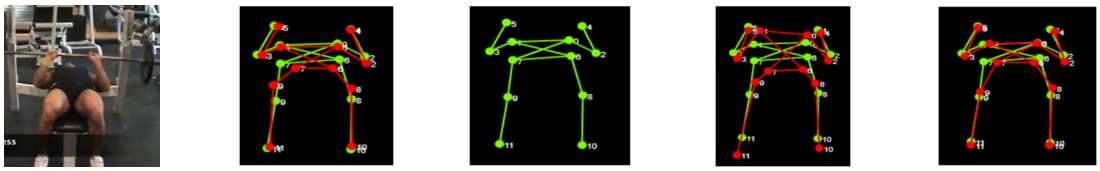



**Figure 7.** Sample video frame and detection result of Action 2 (bench press). Green lines indicate the ground truth, red lines indicate the tested HPE library.

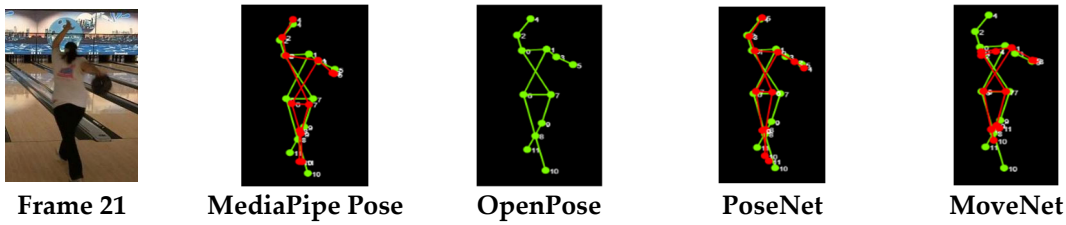



**Figure 8.** Sample video frame and detection result of Action 4 (bowling). Green lines indicate the ground truth, red lines indicate the tested HPE library.

In addition, the PDJ values for each body part in each action were also calculated. Frequent changes in keypoint positions also affects the performance of keypoint detection. In Actions 1, 3, 4, 6, 13, and 14, the PDJ values for elbows and wrists in all the tested libraries were lower than those of other body parts, as shown in Figures 9–11. This was due to frequent changes in keypoint positions in the elbows and wrists compared to other body parts. For instance, the person who plays bowling only needs to use his elbow and wrist to release the ball to the bowling lane, hence showing fewer changes of movement in other body parts. Likewise, the performance of OpenPose (blue line) showed the lowest PDJ among these actions, while the fluctuations in PDJ values between other HPE libraries were relatively small.

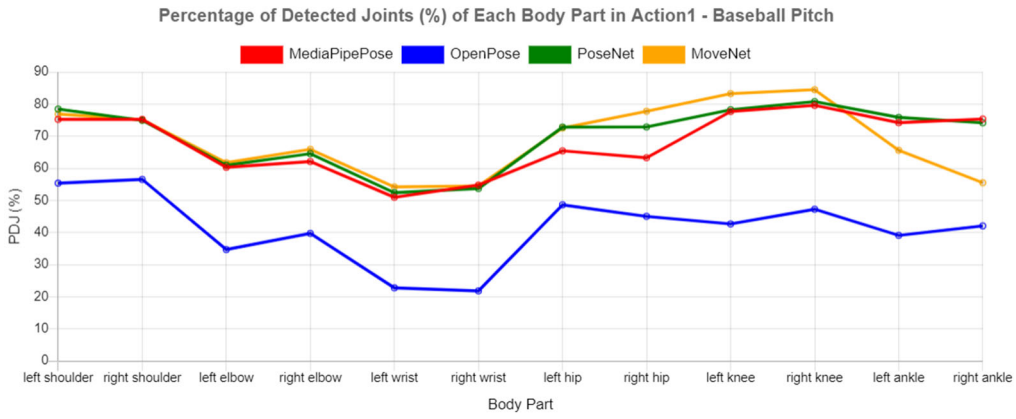



**Figure 9.** PDJ for each body part in Action 1–baseball pitch.

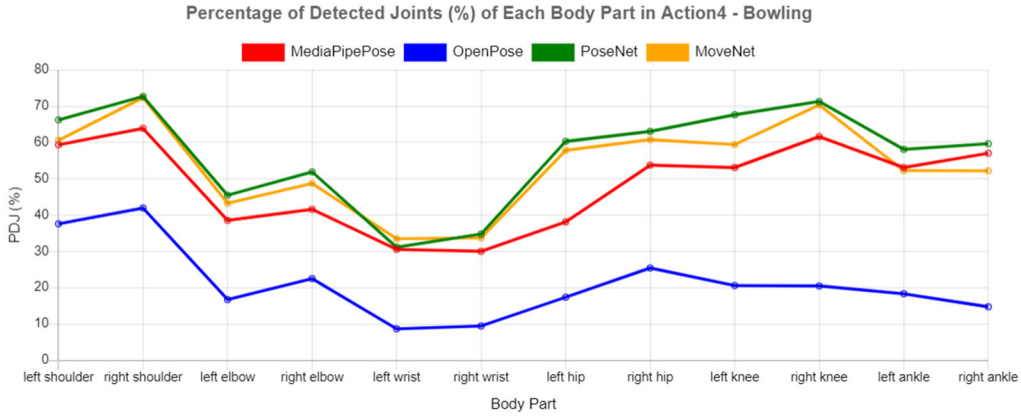



**Figure 10.** PDJ for each body part in Action 4–bowling.

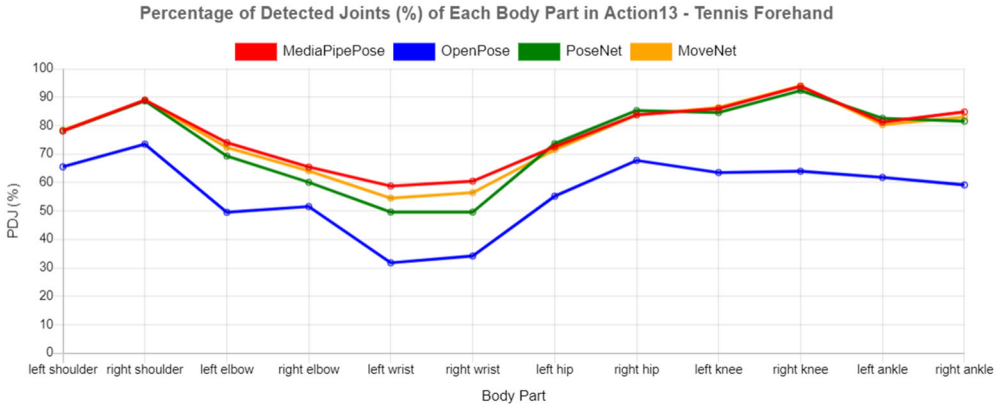



**Figure 11.** PDJ for each body part in Action 13–tennis forehand.

When the common challenges (self-occlusion and inappropriate camera position) occurred in the videos, the performance of keypoint detection was affected in all HPE libraries. Figure 12 shows the video frames of Action 9 (pull up), showing the self-occlusion effect in the action. In this action, all libraries showed much lower performance in some of the body parts compared to that of other body parts due to this self-occlusion effect. The corresponding PDJ values are reported in Figure 13. The PDJ values for ankles (black dotted box) were much lower than other parts because of self-occlusion from the crossed ankles, as shown in Figure 14. OpenPose (blue line) achieved the lowest performance for all body parts, particularly for elbows and wrists. Additionally, the performance of MoveNet for the ankles was lower than that of MediaPipe Pose and PoseNet when facing the self-occlusion effect.

**Frame 1** **Frame 10** **Frame 17**

**Figure 12.** Sample video frames of Action 9 (pull up).

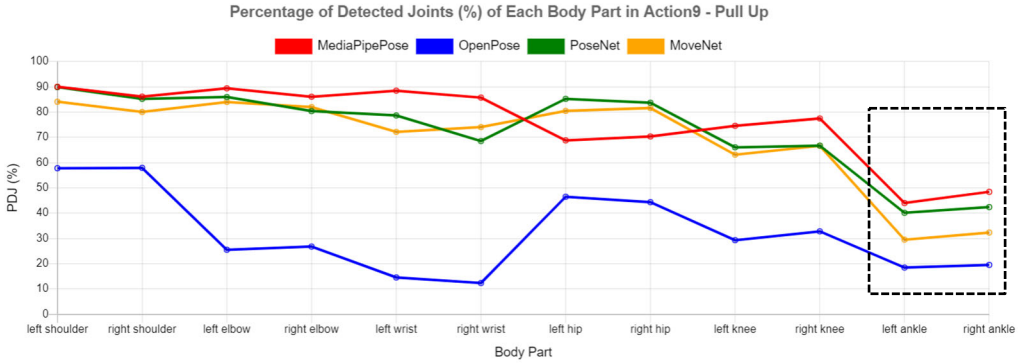
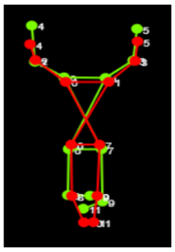
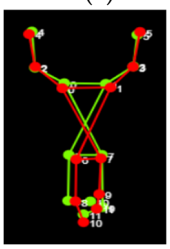



**Figure 13.** PDJ for each body part in Action 9–pull up.

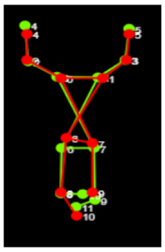
**(a) Video frame 4 of Action 9**



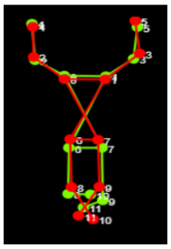
**(b) MoveNet**



**(c) MediaPipe Pose**



**(d) OpenPose**



**(e) PoseNet**

**Figure 14.** (**a**) Frame 4 of Action 9 (pull up). In (**b**–**e**), green lines indicates the ground truth while red lines indicates the tested HPE library.

In addition, inappropriate camera position is one of the common challenges for human pose estimation. In Action 7 (jump rope), the camera was placed on the right side of the person and the video frames were recorded from the right side, as shown in Figure 15. Hence, self-occlusion occurred to the left body parts, which reduced the performance of keypoint detection for the left body parts. Figure 16 clearly shows that the performance of all HPE libraries for the right body parts was higher than that for the left body parts, where the PDJ values for the right shoulder and right elbow were always higher than those for the left shoulder and left elbow.

**Frame 1**

**Frame 7**

**Frame 11**

**Figure 15.** Sample video frames of Action 7 (jump rope).

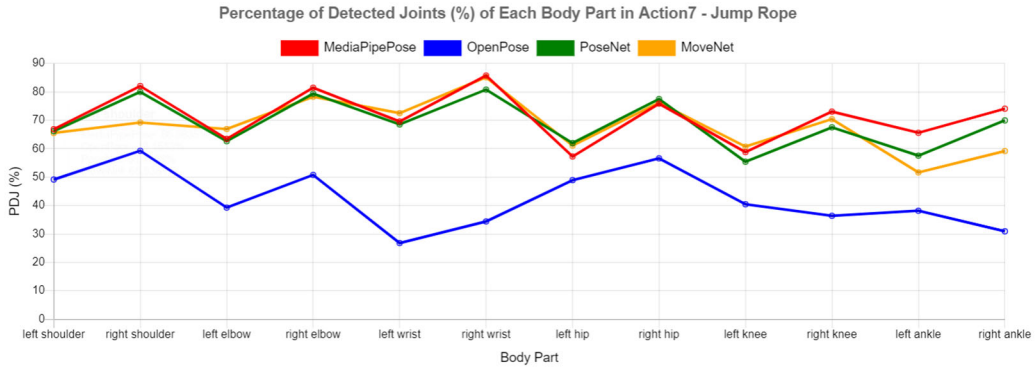



**Figure 16.** PDJ for each body part in Action 7 (jump rope).

In the overall comparison, MoveNet achieved the highest PDJ values in terms of minimum, first quartile, and third quartile values compared to the other HPE libraries. The PDJ values for MoveNet in all actions were more than 50%. MediaPipe Pose achieved the highest PDJ values in terms of maximum and median values. All HPE libraries achieved superior performance in Action 8 (jumping jacks) among all actions, which consists of fewer challenges. Based on the ranking of the HPE libraries in each action, MediaPipe Pose achieved the top performance in 7 actions, followed by MoveNet (6 actions) and PoseNet (1 action). OpenPose showed the poorest performance in all actions. MoveNet achieved the highest average PDJ value, while OpenPose showed the lowest average PDJ value. However, MediaPipe Pose and PoseNet were still competitive with MoveNet because the average PDJ values for these three libraries were in the range of 65–68%.

## 5. Discussion

The main findings of this study are summarized in this section. MediaPipe Pose had the lowest overall performance in the image dataset. However, it performed well in the video dataset with an overall performance slightly lower than the top overall performer, MoveNet. It achieved the best performance in 7 out of 14 actions.

MoveNet achieved the highest overall performance for keypoint detection in the image and video datasets. For the video dataset, the PDJ values for all actions were greater than 50%, which was the best performance compared to the other HPE libraries. Additionally, it achieved the best performance in 6 actions, compared to 7 actions in MediaPipe Pose. However, its overall average PDJ was the highest (69.85%).

OpenPose achieved the second highest performance in the image dataset. However, the weakness of OpenPose was in detecting the keypoints in continuous video frames. Its overall performance in the video dataset was the lowest, approximately 30% lower than the other HPE libraries. Overall, PoseNet had the third highest performance in both the image and video datasets.

The performance of HPE libraries is reduced when they are constrained by challenges such as inappropriate camera position and the self-occlusion effect. OpenPose was the least robust in detecting video frames when facing these challenges. This is because OpenPose always loses track when self-occlusion occurs in the video frames. Figure 17 illustrates this point in the line graph of PDJ values for OpenPose in each frame in Action 1 (baseball pitch). Based on the results, the PDJ values decreased from frame 84 to 102 (refer to the green box). Due to the bottom-up method used by OpenPose, the detected keypoints cannot be grouped into a human instance when there is self-occlusion in these frames, resulting in tracking failure. The original video frames and results of the OpenPose detection from frame 78 to 103 are shown in Figure 18 for illustration. OpenPose obviously performed poorly in frames with self-occlusion (frames 84 to 103).

Based on the results of this experiment, MoveNet is suitable for detecting both images and videos. On the other hand, OpenPose is more suitable for detecting images while

MediaPipe Pose is more suitable for detecting videos. The performance of PoseNet is mediocre.

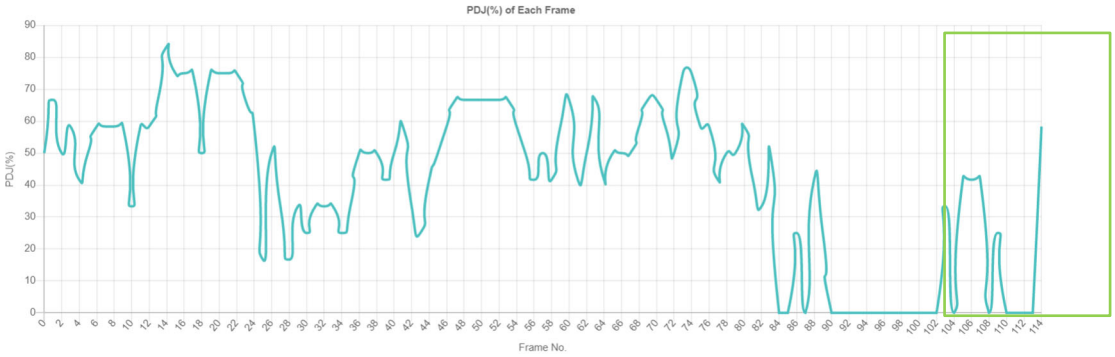



**Figure 17.** Line graph of PDJ values for OpenPose in each frame in Action 1 (baseball pitch).

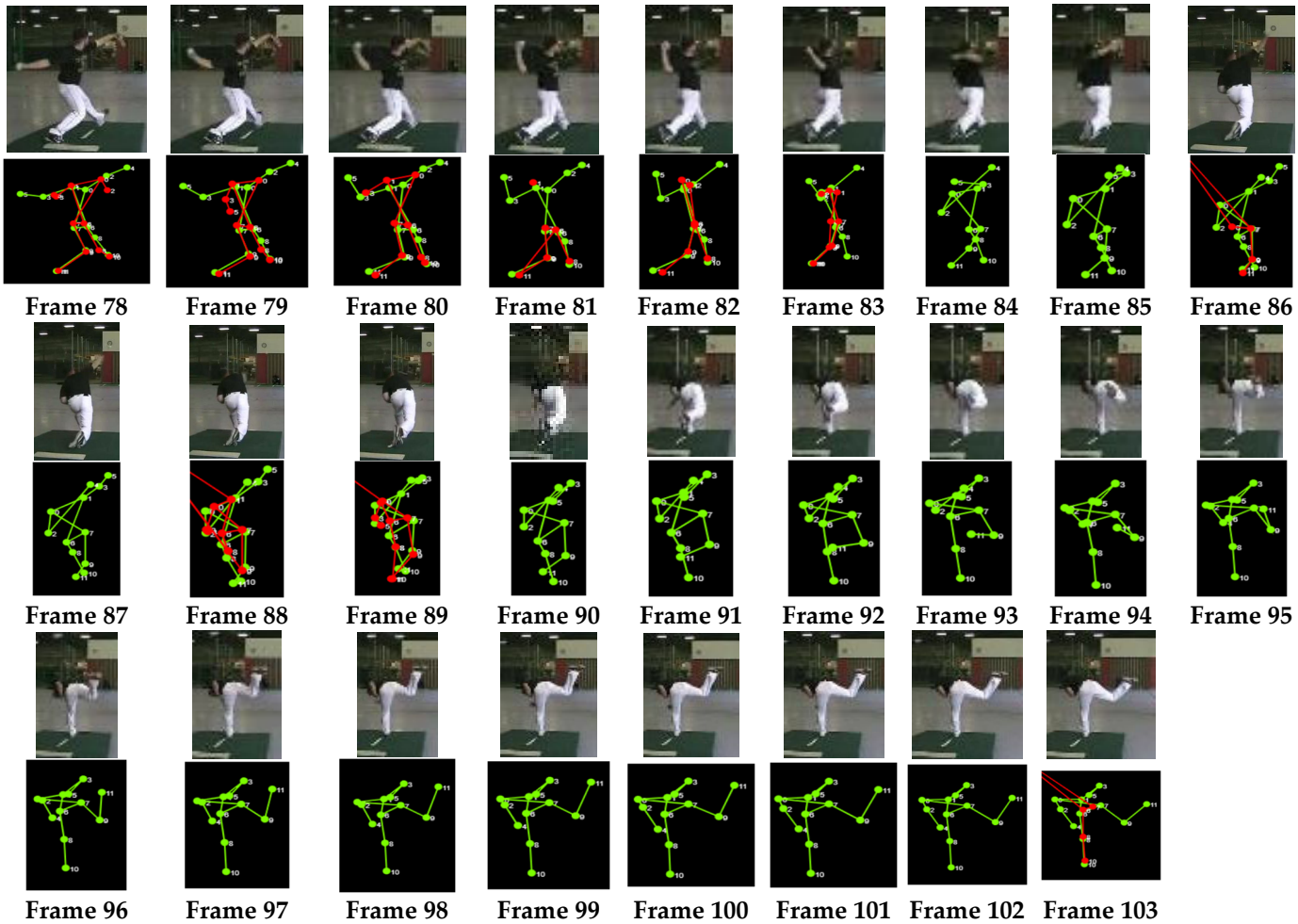



**Figure 18.** Original images and results from OpenPose from frame 78 to frame 103. Green points and lines represent ground truth, while red points and lines represent OpenPose.

## 6. Conclusions

Currently, HPE is a popular task in the field of computer vision because it can be applied in various practical applications. Hence, the performance of HPE libraries is important. This paper has presented a comparative analysis of four state-of-the-art HPE libraries, including OpenPose, PoseNet, MediaPipe Pose, and MoveNet, to investigate their strengths and weaknesses in processing image and video HPE. This study focused only

on 2D single-person HPE. PDJ was the evaluation metric used in this study to evaluate the quality of the HPE libraries.

As a result, MoveNet has superior performance while MediaPipe Pose has the lowest performance in detecting images. In addition, MoveNet also has top performance while OpenPose has the lowest performance in detecting videos. However, OpenPose has the second highest performance in detecting images. PoseNet showed average performance in detecting images and videos. When facing challenges such as inappropriate camera position or self-occlusion, the performance in detecting body parts will be reduced. MoveNet, MediaPipe Pose, and PoseNet can handle these challenges well, but OpenPose shows the poorest performance under these conditions. In detecting videos, OpenPose had the lowest robustness because it loses track when self-occlusion occurs with body parts.

The limitation of this study is that the experiments were focused on analyzing the performance of four HPE libraries using the PDJ. The memory consumption, inference time, and detection speed of each HPE library will be compared in our future work.

**Author Contributions:** Funding acquisition, L.-Y.O.; Investigation, J. -L.C.; Project administration, L.-Y.O.; Supervision, L.-Y.O.; Visualization, J. -L.C.; Writing–original draft, J. -L.C.; Writing–review & editing, L.-Y.O. and M.-C.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Telekom Malaysia Research & Development, RDTC/221036 (MMUE/220003) and Multimedia University IR Fund, MMUI/220001.

**Data Availability Statement:** Not Applicable, the study does not report any data.

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

## References

1. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 29 October 2017; pp. 3960–3969.
2. Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 23 June 2018; pp. 2119–2128.
3. Thyagarajmurthy, A.; Ninad, M.G.; Rakesh, B.G.; Niranjan, S.; Manvi, B. Anomaly detection in surveillance video using pose estimation. In Proceedings of the Emerging Research in Electronics, Computer Science and Technology, 2019; Springer: Singapore, 2019; pp. 753–766. Available online: https://link.springer.com/chapter/10.1007/978-981-13-5802-9_66/ (accessed on 27 October 2022).
4. Lamas, A.; Tabik, S.; Montes, A.C.; Pérez-Hernández, F.; García, J.; Olmos, R.; Herrera, F. Human pose estimation for mitigating false negatives in weapon detection in video-surveillance. *Neurocomputing* **2022**, *489*, 488–503.
5. Yoo, H.R.; Lee, B.H. An openpose-based child abuse decision system using surveillance video. *J. Korea Inst. Inf. Commun. Eng.* **2019**, *23*, 282–290.
6. Park, J.H.; Song, K.; Kim, Y.-S. A Kidnapping Detection Using Human Pose Estimation in Intelligent Video Surveillance Systems. *J. Korea Soc. Comput. Inf.* **2018**, *23*, 9–16. https://doi.org/10.9708/JKSCI.2018.23.08.009
7. Chang, Y.J.; Chen, S.F.; Huang, J.D. A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Res. Dev. Disabil.* **2011**, *32*, 2566–2570.
8. Hassan, H.A.; Abdallah, B.H.; Abdallah, A.A.; Abdel-Aal, R.O.; Numan, R.R.; Darwish, A.K.; El-Behaidy, W.H. Automatic Feedback For Physiotherapy Exercises Based On PoseNet. *FCAI-Inform. Bull*. **2020**, *2*, 10–14.
9. Shapoval, S.; García Zapirain, B.; Mendez Zorrilla, A.; Mugueta-Aguinaga, I. Biofeedback applied to interactive serious games to monitor frailty in an elderly population. *Appl. Sci.* **2021**, *11*, 3502.
10. Chua, J.; Ong, L.Y.; Leow, M.C. Telehealth using PoseNet-based system for in-home rehabilitation. *Future Internet* **2021**, *13*, 173.
11. Kim, W.; Sung, J.; Saakes, D.; Huang, C.; Xiong, S. Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose). *Int. J. Ind. Ergon.* **2021**, *84*, 103164.
12. Jawale, C.D.; Joshi, K.A.; Gogate, S.K.; Badgujar, C. Elcare: Elderly Care With Fall Detection. *J. Phys. Conf. Ser.* **2022**, *2273*, 012019.
13. Kapoor, R.; Jaiswal, A.; Makedon, F. Light-Weight Seated Posture Guidance System with Machine Learning and Computer Vision. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 29 June–1 July 2022; pp. 595–600.

14. Landi, H. Google, ProMedica Team up with IncludeHealth to Tap into Growing Virtual MSK Market. Fierce Healthcare. Available online: https://www.fiercehealthcare.com/tech/google-promedica-team-up-includehealth-to-tap-into-virtual-msk-market (accessed on 16 July 2022).
15. Chen, W.; Jiang, Z.; Guo, H.; Ni, X. Fall detection based on key points of human-skeleton using openpose. *Symmetry* **2020**, *12*, 744.
16. Zou, J.; Li, B.; Wang, L.; Li, Y.; Li, X.; Lei, R.; Sun, S. Intelligent fitness trainer system based on human pose estimation. In Proceedings of the International Conference On Signal And Information Processing, Networking And Computers, Yuzhou, China, 29 November–1 December 2018; Springer: Singapore, 2018; pp. 593–599.
17. Suda, S.; Makino, Y.; Shinoda, H. Prediction of volleyball trajectory using skeletal motions of setter player. In Proceedings of the 10th Augmented Human International Conference, Reims, France, 11–12 March 2019; pp. 1–8.
18. Wang, J.; Qiu, K.; Peng, H.; Fu, J.; Zhu, J. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 374–382.
19. Jeon, H.; Yoon, Y.; Kim, D. Lightweight 2D human pose estimation for fitness coaching system. In Proceedings of the 2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, South Korea, 27–30 June 2021; pp. 1–4.
20. Li, Y.C.; Chang, C.T.; Cheng, C.C.; Huang, Y.L. Baseball Swing Pose Estimation Using OpenPose. In Proceedings of the 2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI), Hong Kong, China, 21–23 April 2021; pp. 6–9.
21. Park, H.J.; Baek, J.W.; Kim, J.H. Imagery based Parametric Classification of Correct and Incorrect Motion for Push-up Counter Using OpenPose. In Proceedings of the 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), Hong Kong, China, 20–21 August 2020; pp. 1389–1394.
22. Nguyen, H.T.P.; Woo, Y.; Huynh, N.N.; Jeong, H. Scoring of Human Body-Balance Ability on Wobble Board Based on the Geometric Solution. *Appl. Sci.* **2022**, *12*, 5967.
23. Patil, A.; Rao, D.; Utturwar, K.; Shelke, T.; Sarda, E. Body Posture Detection and Motion Tracking using AI for Medical Exercises and Recommendation System. *ITM Web Conf.* **2022**, *44*, 03043
24. Devanandan, M.; Rasaratnam, V.; Anbalagan, M.K.; Asokan, N.; Panchendrarajan, R.; Tharmaseelan, J. Cricket Shot Image Classification Using Random Forest. In Proceedings of the 2021 3rd International Conference on Advancements in Computing (ICAC), Colombo, Sri Lanka, 9–11 December 2021; pp. 425–430.
25. Joseph, R.; Ayyappan, M.; Shetty, T.; Gaonkar, G.; Nagpal, A. BeFit—A Real-Time Workout Analyzer. In Proceedings of the Sentimental Analysis and Deep Learning, Springer: Singapore, 2022; pp. 303–318. Available online: https://link.springer.com/chapter/10.1007/978-981-16-5157-1_24/ (accessed on 27 October 2022).
26. Mahendran, N. Deep Learning for Fitness. *arXiv* **2021**, arXiv:2109.01376.
27. Agarwal, S.; Gupta, M.; Khandelwal, S.; Jain, P.; Aggarwal, A.; Singh, D.; Mishra, V.K. FitMe: A Fitness Application for Accurate Pose Estimation Using Deep Learning. In Proceedings of the 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC), Jalandhar India, 21–23 May 2021; pp. 232–237.
28. Nakai, M.; Tsunoda, Y.; Hayashi, H.; Murakoshi, H. Prediction of basketball free throw shooting by openpose. In Proceedings of the JSAI International Symposium on Artificial Intelligence, Yokohama, Japan, 12–14 November 2018; Springer: Cham, Switzerland, 2018; pp. 435–446.
29. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. (). Deep learning-based human pose estimation: A survey. *arXiv* **2020**, arXiv:2012.13392.
30. Papandreou, G.; Zhu, T.; Kanazawa, N.; Toshev, A.; Tompson, J.; Bregler, C.; Murphy, K. Towards accurate multi-person pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4903–4911.
31. Jo, B.; Kim, S. Comparative Analysis of OpenPose, PoseNet, and MoveNet Models for Pose Estimation in Mobile Devices. *Trait. du Signal* **2022**, *39*, 119–124.
32. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
33. Bazarevsky, V.; Grishchenko, I.; Raveendran, K.; Zhu, T.; Zhang, F.; Grundmann, M. BlazePose: On-device Real-time Body Pose tracking. *arXiv* **2020**, arXiv:2006.10204.
34. Gadhiya, R.; Kalani, N. Analysis of Deep Learning Based Pose Estimation Techniques for Locating Landmarks on Human Body Parts. In Proceedings of the 2021 International Conference on Circuits, Controls and Communications (CCUBE), December 2021; pp. 1–4. Available online: https://ieeexplore.ieee.org/abstract/document/9702726/ (accessed on 27 October 2022).
35. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
36. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 28 June 2014; pp. 3686–3693.

37. 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, 28 June 2014; pp. 1653–1660.
38. Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P.V.; Schiele, B. Deepcut: Joint subset partition and labeling for multi person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 30 June 2016; pp. 4929–4937.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
40. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
41. 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.* **2019**, *43*, 172–186.
42. Zhou, X.; Wang, D.; Krähenbühl, P. (2019). Objects as points. arXiv preprint arXiv:1904.07850.
43. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
44. Zhang, W.; Zhu, M.; Derpanis, K.G. From actemes to action: A strongly-supervised representation for detailed action understanding. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2248–2255.
45. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
46. Varamesh, A.; Tuytelaars, T. Mixture dense regression for object detection and human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13086–13095.
47. Wang, M.; Tighe, J.; Modolo, D. Combining detection and tracking for human pose estimation in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11088–11096.
48. Artacho, B.; Savakis, A. Unipose: Unified human pose estimation in single images and videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7035–7044.
49. Luvizon, D.C.; Picard, D.; Tabia, H. 2d/3d pose estimation and action recognition using multitask deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA 18–23 June 2018; pp. 5137–5146.
50. Liu, J.; Shi, M.; Chen, Q.; Fu, H.; Tai, C.L. Normalized human pose features for human action video alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11521–11531.
51. Ahmedt-Aristizabal, D.; Nguyen, K.; Denman, S.; Sridharan, S.; Dionisio, S.; Fookes, C. Deep motion analysis for epileptic seizure classification. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 3578–3581.
52. You, Y.; Zhao, Y. A human pose estimation algorithm based on the integration of improved convolutional neural networks and multi-level graph structure constrained model. *Pers. Ubiquitous Comput.* **2019**, *23*, 607–616.