総説

Potential Use of Deep Learning-Based Pose Estimation in Sports Biomechanics
ディープラーニングを用いた骨格推定のスポーツバイオメカニクスでの活用可能性

Hiroki Ozaki[1], Minoru Matsumoto[1], Hideyuki Nagao[1),2], Toshiharu Yokozawa[1]
尾崎宏樹 [1], 松本実 [1], 長尾秀行 [1),2], 横澤俊治 [1]

Abstract : Pose estimation using deep learning (DL) is expected to solve traditional problems faced by sports biomechanics, including limitations resulting from the application of reflective markers. For sports biomechanists to correctly utilize these pose estimation techniques, there is a need to elucidate the estimation and learning procedures used in pose estimation as well as to consider how to utilize them. Therefore, we aimed to review recently published major pose estimation models and to examine the availability of pose estimation in sports biomechanics. We observed that the main models were developed for simultaneous estimation of multiple persons, but none of the these were designed to rigorously estimate center of joint position which is mainly required in sports biomechanics. Further, all training datasets for these models were digitized positions that appeared as the joint centers of people in "in-the-wild" videos; moreover, these workers were non-professionals termed as "crowd-workers". Therefore, regardless of the model quality, the dataset accuracy may be a bottleneck that impedes the estimation accuracy required in sports biomechanics. All the metrics used to verify the accuracy involved verification of the average estimation results of multiple joint points across the entire frame or multiple frames. Therefore, even with a high overall estimation accuracy, the accuracy of the estimated positions of the individual joints may be low. Taken together, it is difficult to utilize and calculate kinematic variables based on joint positions obtained through pose estimation. However, the existing pose estimation may help sports biomechanists calculate the movement periodization and number. To expand the utility of pose estimation in sports biomechanics, sports biomechanists should be actively involved in the development of pose estimation models and datasets.

Key words : dataset, verification metrics, deep learning, marker-less, motion capture.
キーワード：データセット, 評価指標 , ディープラーニング, マーカレス, モーションキャプチャ

[1]Japan Institute of Sports Sciences, [2]Sakushin Gakuin University

[1] 国立スポーツ科学センター , [2] 作新学院大学

E-mail : hiroki.ozaki@jpnsport.go.jp

受付日：2022 年 3 月 23 日

受理日：2022 年 10 月 7 日

Ⅰ. Introduction

1. Athletes and sports biomechanics

Athletes require technical and physical training to enhance their athletic performance. Sports biomechanists examine the factors that limit performance enhancement among athletes and suggest means of improving appropriate motions. To identify factors limiting athletic performance, sports biomechanists analyze the athlete's athletic performance (hereafter referred to as performance analysis) or motion (hereafter referred to as motion analysis). In the performance analysis, they examine key variables, including the distance traveled per section and the travel speed, to identify the factors limiting athletic performance. In the motion analysis, several fundamental motions, including running, jumping, and throwing, as well as competition-specific motions that influence athletic performance, may also be analyzed to examine limiting factors in each motion. Sports biomechanists generally begin the analysis process by considering each body part as a rigid segment. Accordingly, the initial step is creating the athlete's body model, with rigid body segments being linked by joints (the so-called link-segment model). In this initial step, biomechanists must use accurate joint position data. This is because slight differences in the joint center position can significantly affect biomechanical variables, including segment velocity, joint force, and moment, which are calculated based on the joint center position. For example, Reinschmidt et al. compared bone pins and skin markers and showed that measurement errors occurring in hip adduction abduction and internal rotation external rotation affect the process of motion improvement [40]. In addition, Leardini et al mentioned that a hip joint center misplacement of 30 mm in the anterior–posterior direction generated a mean error on the flexion/extension moment of about 22%[27]. This error also has a significant impact when making suggestions to improve movements. Moreover, those kinds of errors can be a problem when comparing the kinetics data from different test conditions or those presented by different research laboratories which use different alignment procedures[4]. Therefore, sports biomechanists typically first mark anatomical landmarks (bone ends near a joint of interest whose position can be reproducibly identified from the skin) to accurately estimate the joint center position by palpation[51]. The required anatomical landmarks are marked using reflective spherical or semi-spherical markers; moreover, their 2D or 3D positions are used as cues to calculate the joint center positions.

2. Limitations in performance and motion analysis methodologies.

The aforementioned method is a powerful process for accurately determining the joint center position; however, it has several limiting factors. For example, application of reflective markers is difficult during sporting activities[35]. Additionally, placing markers on the body areas sensitive to manipulation, including fingertips in baseball pitching, may interfere with the athlete's motion. Generally, the digitization method (cricking the joint center position in a video clip using specific software and providing information as the joint center position) and the Mo-cap method (identifying reflective marker positions using a 3D-motion analysis system) are used to calculate the positions of reflective markers. However, these methods also have several limitations. The digitization method involves risks such as human errors in identifying the marker position due to manual digitization. Additionally, digitization requires a significant amount of time. The Mo-cap method also has limitations, including the expensive Mo-cap system and the fact that it can only analyze motions within a limited space. Motion capture using IMUs can solve some of these problems. However, this method also has limitations, including vibration of the sensor itself and data drift. Therefore, this would not be a substitute for digitization or Mo-cap[15].

3. Accuracy required for pose estimation using deep learning in sports biomechanics.

Recently, there has been rapid development of deep learning (hereinafter referred to as DL) in various fields. This is especially true in the image processing field, where new technologies such as security systems based on face recognition and medical assistance systems based on X-ray images have been created. In sports biomechanics, there have been attempts to utilize DL, including a table tennis rally detection algorithm[9] and ball-tracking technologies[16), 41]. Additionally, another notable technique is pose estimation. Since the development of DeepPose[47], various DL-based pose estimation methods have been described. These DL-based pose estimations could solve the limitations impeding joint center position identification by sports biomechanists. In sports biomechanics, there are various ways of accurately identifying the joint center position. In the imaging method, great care is taken to ensure the following: reflective markers attached to the individual's anatomical landmarks appear on the image with high resolution, the camera is positioned to extensively minimize occlusion of reflective markers, and the appropriate imaging speed is selected to match the movement speed especially when capturing fast and dynamic movements[51]. Further, trained biomechanists usually digitize the center position of the reflective markers in the image and multiple people can digitize the reflective markers within the same trial and average the results. Another idea is suppressing the vibration of reflective markers caused by skin movement. The skin is shaken when the target motion involves impact, including sprinting or hitting, with this effect being mixed into the joint center position information as noise. There are various measures proposed to at least minimally reduce this influence, including the use of cluster markers[7] or the symmetrical center of the segment's rotation[34]. On the other hand, existing models and datasets are considered to be designed for pose estimation in daily life. Therefore, it is unclear whether existing models can adequately estimate high-speed, acrobatic movements and rotational movements around the long axis, such as the supination and pronation of the forearm seen in baseball and other sports. In addition, while sports biomechanics requires accurate positions of target joints in order to suggest improvements in movement, the development of pose estimation to date is thought to be envisioned as a contribution to computer industries including video game annotation[11]. Therefore, for pose estimation to be applied in sports biomechanics, it is necessary to determine how to accurately obtain joint positions, and whether these accurately meets the standards in sports biomechanists.

4. Study objectives

In this context, to consider the use of DL-based pose estimation in sports biomechanics, there is a need to determine whether the method of obtaining the joint center position in the pose estimation model using DL allows the appropriate accuracy required in sports biomechanics. The accuracy of the joint center position information contained in the dataset used as supervisory data significantly affects the accuracy of joint center position estimation through pose estimation[43]. There have been a number of review papers on pose estimation[4), 45), 49]. These papers summarized the DL-based pose estimation models, the datasets of those models, and the accuracy validation of the last decade, but they did not address the potential uses of pose estimation in sports biomechanics. Colyer et al. conducted a narrative review of DL-based pose estimation models from a sports biomechanics perspective[11]. However, they mainly discussed 3D pose estimation without markers and did not review their datasets. Although many reviews have been published, it still remains unclear whether the methods currently used to validate the accuracy of 2D, and 3D pose estimation models that can be used for sports biomechanists. Therefore, the purpose of this study was to review recent posture estimation models, datasets, and their accuracy validation to determine their applicability to posture estimation, particularly in sports biomechanics.

Ⅱ. Literature Review Procedures

We conducted a comprehensive literature review on (1) DL-based pose estimation models that may be relevant to sports biomechanics, as well as (2) the applied teacher datasets and (3) accuracy evaluation metrics in this study. For (1), papers accepted to the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, the top conference for human pose estimaton[49] that were published after 2014 and cited at least 50 times according to Google Scholar as of June 1, 2022, were included in this study. However, since the most recent papers may have been excluded due to citation counts, this review also included the most recent review papers published after 2019 that were presented as models incorporating new methods. Therefore, (2) Teacher datasets and (3) accuracy evaluation were covered by investigating citations used in studies selected in (1).

1. Definition of terms

The words used in this article are defined as follows:

Pose estimation: A technique for pose estimation of a person or animal based on images or videos. Although techniques for pose estimation have been developed since before the introduction of machine learning techniques, we defined them as techniques using DL.

Model: In this review, this term is synonymous with machine learning models. The model is a concrete calculation process that, upon reception of input, evaluates the input's content produces an output value.

Dataset: A set of data collected for a specific purpose. In this review, this term specifically refers to a collection of data in which images or videos, which serve as supervisory data in DL, and labels are collected as a set.

2. Recent trends in pose estimation models, datasets, and accuracy verification methods

1) Pose estimation models

a) Recent trends in pose estimation models

Emergence of DeepPose

Since the early 2010s, there has been rapid development of DL given the spread of the Internet, the reduced computation time by general-purpose GPU, and the development of important algorithms such as convolutional neural networks (CNNs). After the development of DeepPose, various pose estimation models using DL have been proposed (Table 1).

Trends in the development of pose estimation models after DeepPose

DeepPose[47] was the first pose estimation model using DL, but it was only slightly more accurate than non-DL-based models. Tompson et al.[46] reported that heat maps could effectively improve the estimation accuracy of pose estimation. They greatly improved the accuracy of joint position estimation from DeepPose using heat maps. Since then, heat-map-based models have been proposed. Other methods for improving accuracy using heat maps in stages have been developed[50), 36], with further improvement in the accuracy of the pose estimation. These achievements have allowed the simultaneous detection of both single and multiple persons. DeepCut[37] and DeeperCut[18] were developed to initially detect only a person, followed by estimation of the individual's joints from a video showing multiple persons, which improved the speed of the pose estimation process for multiple persons. DeepCut has improved the method for estimating the whole-body skeleton from joint positions, and has improved the estimation accuracy of the lower body in particular from the method of Tompson et al[46].

DeepCut also achieves highly accurate joint position estimation in a short computation time. In these models, information regarding the pose estimation was not passed between frames; therefore, in case of mul-

Table 1: Main Pose estimation models using deep learning and their overview. Note that NN indicates that the model is not specifically named in the paper. Listed from oldest to most recent.

Model	Multi person	Tracking	3D	Dataset
DeepPose[47)]	-	-	-	FLIC, LSP, LSP Extended
NN[46)]	-	-	-	FLIC, LSP Extended
Convolutional Pose Machines[50)]	-	-	-	FLIC, LSP, MPII
NN[36)]	-	-	-	FLIC, MPII
DeepCut[37)]	✓	-	-	LSP, LSP Extended, MPII
DeeperCut[18)]	✓	-	-	LSP, LSP Extended, MPII
OpenPose[8)]	✓	-	-	MPII Multi-person, MS COCO keypoints
ArtTrack[17)]	✓	✓	-	MPII Multi-person
RMPE[12)]	✓	-	-	MPII Multi-person, MS COCO keypoints
NN[10)]	-	-	✓	LSP, Human3.6M
PoseFlow[52)]	✓	✓	-	MPII Multi-person, PoseTrack
DensePose[14)]	✓	-	✓	MS COCO keypoints, Densepose COCO
Human Mesh Recovery[24)]	-	-	✓	LSP, LSP-extended, MPII, MS COCO keypoints, Human3.6M, MPI- INF-3DHP
CrowdPose[28)]	✓	-	-	MS COCO keypoints, CrowdPose, J-HMDB
NN[39)]	✓	✓	-	MPII, MS COCO keypoints, PoseTrack
NN[53)]	-	-	✓	Human3.6M, HumanEva
NN[20)]	✓	-	✓	Human3.6M, LSP, LSP-extended, MPII, MPI- INF-3DHP, MS COCO and more
NN[29)]	✓	-	✓	Human3.6M, MPI- INF-3DHP and more
VIBE[25)]	✓	-	✓	Human3.6M, MPI- INF-3DHP

tiple detected persons, they would be swapped every time the frame changed. To address this issue, several methods have been proposed to simultaneously estimate and track multiple persons[12), 17)]. These methods not only improved the accuracy of pose estimation for multiple persons, but also improved the accuracy of joint estimation for a single person. For example, the RMPE proposed by Fang et al[12)] outperformed the estimation accuracy of DeeperCut in head, shoulder, elbow, wrist, hip, knee and ankle segments. In these proposed models, the chin position, which is relatively easy for the model to identify, is first estimated, followed by estimation of the joint center position of the whole body is estimated from the positional information (top-down method). Subsequently, these models can quickly detect and track the person. However, there remained issues with the top-down method, including the fact that pose estimation cannot be performed upon failure of the recognition of a person and the misidentification of other persons' joints as those of the target person. Another limitation is the computational cost, which is positively correlated with the number of people. To address these challenges, Xiu et al.[52)] designed a method for reducing non-skeletal noise in images known as PoseFlow that addressed fast motion, occlusion, and motion blurring using cross-frame detection. Li et al. also devised a model that first identifies and detects the range of the person within the image; subsequently, it considers the person to be the same even if body parts are outside the range (CrowdPose[28)]). These models improved the accuracy of pose estimation of densely packed people in sports, which was previously difficult to estimate. Here, they adopted a method of estimating joint center positions from the entire image (the so-called bottom-up method). By accounting for the orientation from one joint to another adjacent joint, they solved the problem of segments not being connected well, which was a limitation of the conventional bottom-up method, and solved the problem of the aforementioned top-down method. Besides these, there are some models[8), 39)] that allow highly accurate multi-person pose estimation. Since 2017, multiple 3D pose estimation algorithms have been reported in the computer science field[10), 24), 53)]. Numerous computer scientists have proposed new innovations for pose estimation in 3D include estimating 3D coordinates of joint points through regression with image input to CNN, retrieving 3D postures corresponding to 2D postures from a library, and fitting a human model. Güler et al.[14)] proposed a novel method for mapping pixels of the human body in an image to a 3D human surface. However, it only qualitatively verified that the posture estimation by this method was highly accurate regardless of the presence of multiple persons, postures, costumes, scales, and occlusions, and did not compare the accuracy of this method with other models. The 3D pose estimation created a new challenge. Specifically, the overlap of people, depth perception, and anterior-posterior relationships could not be correctly estimated, which were not as problematic in 2D pose estimation. To address this issue, there have been numerous recent reports regarding a technique called human shape reconstruction, which solves the problem of inconsistencies in the anterior-posterior relationships of body parts when identifying multiple overlapping people and estimating their skeletons. Jiang et al.[21)] proposed a model that accounts for the correct estimation of the anterior-posterior relationships of multiple people detected in 2D images (depth ordering-aware loss), which improved the overlap problem in 3D pose estimation. Their complete model allowed them to generate more coherent reconstructions. However, the improvement in estimation accuracy relative to previous studies was marginal. Research in this area is expected to accelerate as pose estimation becomes more 3D. One of the problems with 3D pose estimation was the ambiguity mainly caused by occlusion when projecting a 2D pose into 3D[23), 33)]. To reduce this ambiguity, multi-view images and video sequences input were proposed. Liang et al. [29)] improved the accuracy in joint position estimation by 3D pose estimation assuming Virtual Try On for both single and multi-person by simultaneously inputting images from three to four different orientations. While many attempts at video input have been reported, Kocabas et al. [25)] specifically improved the accuracy of posture estimation for a single person by propagating posture information over time and introducing self-attention in the discriminator to focus on the important temporal structure of human motion.

b) Summary of trends in pose estimation models

Since the development of DeepPose, the accuracy of pose estimation has been improved by innovative ideas such as heat maps, bottom-up methods, bottom-down methods, and cross-frame detection. Accordingly, the performance of pose estimation models has been improved through simultaneous estimation of multiple persons. Additionally, pose estimation in 3D has become popular. In the 3D pose estimation, multi-view image and video sequence input contributed to the improvement of 3D pose estimation accuracy in order to reduce the ambiguity that had been an issue when projecting 2D poses into 3D. However, although 3D methods contribute to multi-person identification, they do not significantly improve the accuracy of joint center position estimation. Further, there have been no reports regarding the development of the models that were designed to rigorously estimate center of joint position. This indicates that the development motivation of computer scientists does not necessarily match the needs of sports biomechanists.

2) Datasets

Since the early 2000s, there have been several reports on datasets with names (labels) of objects and actions in images and videos. Among the factors that facilitated the rapid development of pose estimation was the progress in the development of teacher datasets. In this section, we review papers on the datasets used in the models reviewed in Section Ⅱ-2-1) (Table2).

a) Datasets containing sports actions

The Leeds Sports Pose Dataset (LSP[22]) is a dataset containing whole-body joint center position information, including the head, in 2,000 images. It digitizes the joint center positions in images collected from Flicker social media. The specific digitization method was not specified. The dataset included sports images of athletes in track and field, badminton, baseball, soccer, tennis, and volleyball. The Leeds Sports Pose Extended Training Dataset (LSP Extended[23]), which was released the following year, added images of athletes in parkour, gymnastics, and other activities not included in the LSP. Digitization was performed using a part-time work-sharing system provided by Amazon Mechanical Turk (AMT[1]). Frames Labeled in Cinema (FLIC[42]) disclosed the positional information of the left and right shoulders, elbows, wrists, hips, knees, and ankles of movie images and the persons in it, with digitization using the AMT. The Joint-annotated Human Motion Data Base (J-HMDB)[20] contains 21 movements, 3,838 frame images, and digitized joint

Table 2: The datasets used for Pose estimation models reviewed in this study and its characteristics. Listed from oldest to most recent.

Dataset	Sports	Multi Person	3D	Key Points	Data size	Source	Annotation
LSP[22]	✓	-	-	full body Joints, head	2,000 annotated images	Flicker	Unknown
HumanEva[44]	-	-	✓	full body Joints, head	over 80,000 calibrated images, 4 persons, 6 actions	-	Motion capture
LSP Extended[23]	✓	-	-	full body joints, head	10,000 annotated images	LSP, Flicker	AMT
FLIC[42]	-	-	-	upper body joints	5,003 annotated images	popular Movies	AMT
J-HMDB[20]	✓	-	-	full body joints, face, belly	928 videos, 33,183 annotated images	HMDB51[26], Internet	AMT
Human3.6M[19]	-	-	✓	full body Joints, head	3.6 million 3D annotated images, 11 persons, 17 actions	-	Motion capture
MS COCO keypoints [30]	-	✓	-	full body joints, eyes, nose, ears	200,000 images, 250,000 persons	Flicker	AMT
MPII[3]	-	✓	✓	full body joints, eyes, nose	25,000 images, over 40,000 persons	YouTube	AMT
MPI-INF-3DHP[33]	-	-	✓	full body Joints, head	over 1.3 million 3D annotated images, 8 persons, 8 actions	-	AMT
PoseTrack[2]	-	✓	-	full body joints, head, nose, neck	1,356 videos, 46,000 annotated images	MP II	Motion capture
DensePose-COCO[14]	-	✓	✓	full body Joints, head	more than 5 million annotated correspondences, 50,000 persons	MS COCO keypoints	Playment

center positions collected from the Internet, including golf swings, ball kicks, and baseball swings.

b) 3D datasets

There are several published datasets for 3D pose estimation. HumanEva[44] published a dataset using VICON (VICON Inc., UK), which is a motion capture system commonly used for motion analysis in sports biomechanics, as a gold standard for quantitative evaluation of pose estimation models. The dataset contains seven videos of four individuals performing actions, including walking, jogging, and gesturing. The position where the markers are attached is chosen to create a human figure using stick pictures rather than identifying the joint center. Moreover, Human3.6M[19] used a motion capture system to capture 11 professional actors within ≈ 3.6 million video frames. A total of 17 scenarios (discussion, smoking, taking a photo, talking on the phone, etc.) were performed in a laboratory environment. Since they were measured using a motion capture system, the 3D marker positions are considered accurate. However, it seems that they were not intended to calculate joint center positions with the accuracy required in sports biomechanics because the markers were attached to ruse clothing where the joints were not visible.

c) Other datasets

Microsoft Common Objects in Context (MS COCO keypoints[30]) is a large dataset of digitized eyes, noses, ears, and joints of the whole body, with additional object detection, segmentation, and image captioning. It is comprised of 330,000 images, with five captions per image (thing categories such as a person, bicycle, and elephant) and a subset of 91 stuff categories (grass, sky, and road). Furthermore, sports scenes are included; however, rather than categorizing sports, the dataset includes sports scenes as among the landscapes. It does not describe the image sources or digitization method. The MPII Human Pose Dataset (MPII[3]) also contains skeletal data of more than 40,000 people with 25,000 database images of joints of the whole body, eyes, and noses, with YouTube as a source. Digitization was performed using AMT. PoseTrack[2] involves more detailed digitization of the MPII Human Pose Dataset and organizes the dataset as a new benchmark for evaluating pose estimation models. Over 66,000 images from 550 videos were extracted and organized for training, validation, and testing. Digitization was performed using an online service similar to AMT called Playment[6].

d) Summary of dataset features

Taken together, datasets that include sports images as a single category include LSP, LSP Extended, and J-HMDB; contrastingly, the other datasets only include scenarios in which people happen to be participating in sports. Most datasets were digitized from “in-the-wild” images that are publicly available online, including social media. This is a reasonable solution since there is no need to re-film the images; however, since the images were not filmed for digitization, occlusion or low resolution is not considered to improve the digitization accuracy. Two articles did not mention the digitization method; however, most of them used part-time workshares from AMT and Playment. Dataset producers can use “crowd-workers” registered on these workshare sites to perform digitization. Crowd-workers are paid based on the number of working man-hours. These sites can have requirements for selecting individuals who can handle the task, for example, a master’s degree. At the same site, the dataset producer can provide special instructions for workers. Indeed, Andriluka et al.[3] stated that “We pre-select AMT workers based on a qualification task and then maintain data quality by manually inspecting the annotated data.” Care was taken to maintain the digitization accuracy. However, the articles did not describe any confirmation of the digitization results. Additionally, given the images published in each dataset, there were numerous cases where the digitization of the joint center position was performed on clothes or on a point clearly off the joint center. In the HUMANEVA and Human3.6M datasets, which include 3D joint center positions, a motion capture system was used to calculate the joint center positon and take. However, there was a low accuracy of joint position estimation since reflective markers were applied to the clothes; moreover, it is unsuitable as a dataset for accurate estimation of joint positions in sports activities since it only deals with movements that can be performed within the range where the motion capture system is installed. Finally, since all 3D datasets were taken in the lab, there was a lack of diversity of the background environment, appearance of the person's clothes, and their postures[11].

3) Verification method of pose estimation models

a) Evaluation indices used for verification

This section summarizes the characteristics of the indicators for accuracy verification of the aforementioned pose estimation models and datasets. The accuracy verification methods included Object Keypoint Similarity (OKS), Mean Average Precision (mAP), Percentage of Correct Part (PCP), Percentage of Correct Key-points (PCK), Multiple Object Tracking (MOTA), and Mean Per Joint Position Error (MPJPE).

Object Keypoint Similarity[13]

It represents the average similarity between the estimated and correct coordinates for the digitized joint center position. This index includes the visibility of the joint center positions (unlabeled, labeled but not visible, labeled and visible) and the evaluation criteria (keypoint similarity) regarding the distance between the correct and estimated positions of each joint center position. The OKS uses the average of multiple keypoint similarities over the entire person. However, it is unsuitable for evaluating the entire algorithm in case there are multiple persons.

Mean Average Precision[54]

The mAP is a measure of the degree to which the boxes of all objects to be detected overlap with the expected boxes. It is sometimes referred to as OKS-based mAP since it is based on OKS. mAP score is calculated by taking the mean average precision over all classes and/or overall IoU (intersection over union) thresholds.

Percentage of Correct Part[3]

The PCP is an index that determines whether a body segment has been detected and evaluates the detection rate. Detection is judged based on whether the length of the line connecting both ends of the estimated segment is within the allowable range for the correct segment length. Since the segment length is used as a criterion, if the segment length in the video is short, the number of pixels that comprise a segment is reduced. Therefore, the effect of a small error becomes relatively larger and the detection criterion also becomes relatively strict. To address this problem, the average of all segment lengths can be used as the standard for evaluation.

Percentage of Correct Key-points[54]

The PCK judges the estimated joint center position as correct when the distance between the estimated and correct coordinates of the joint center position is within the allowable range. The PCK is an index where the evaluation value is the percentage of correct estimation. There are variations according to the application, including using the PCKh when the threshold is determined based on the head size (length of the diagonal line of the bounding rectangle of the head).

Multiple Object Tracking[5]

The MOTA is a measure of the accuracy of object tracking of key points within multiple frames. It evaluates the accuracy of object recognition using the total number of false and missed positives divided by the number of recognized objects. The feature of this indicator is that the number of times that the IDs waved on the objects are accidentally switched when they intersect is accounted for in the evaluation.

Mean Per Joint Position Error[19)]

The MPJPE is used to verify the accuracy of the pose estimation in 3D. It is calculated by averaging the distance between the estimated and correct coordinates across all joint points and data. In the case of algorithms using a single camera, the estimated and correct poses are sometimes done by performing rigid alignment, which involves translating or rotating the estimated pose before evaluation. Therefore, for between-study comparisons of evaluation values, care must be taken to ensure that there are no differences in the evaluation procedure.

b) Summary of the evaluation metrics used for verification

All six aforementioned evaluation metrics were used to determine the error between the joint center position estimated by the pose estimation model and the correct data provided by the dataset. All these metrics verified the accuracy of the entire image or the average of multiple images. Therefore, it is an excellent index for verifying the accuracy of the entire image or video and the average accuracy of the pose estimation for an entire person. Additionally, it is an effective index for discussing the accuracy of different models since it allows objective model evaluation using a common dataset and evaluation index. However, these evaluation metrics cannot individually verify the estimation accuracy of the individual joint center positions in a person, or a certain object point at a certain moment, which is a minimum requirement in sports biomechanics. Vafadar et al.[48)] calculated the segment length based on the positions of the elbow, wrist, hip, knee, and ankle joints and compared CNN-based human pose estimation and marker-based motion analysis system. They also verified the accuracy of each joint by using Mean absolute error (MAE). Using MAE allowed them to determine which joint positions affected the segment length calculation results. Vafader et al. stated that the accuracy was often exacerbated when projecting a 2D pose into 3D[48)]. In this case, the first step might be to carefully conduct accuracy verification in 2D. The MAE was not used in the model papers in this review but it is possible that it was overlooked due to the method used to collect the papers. However, the MAE is considered to be useful not only in 3D but also in 2D in sports biomechanics, because it at least enables verification of the accuracy of each joint.

Ⅲ. Potential applications in sports biomechanics

1 . Current status and issues of pose estimation models, datasets, and their evaluation indices

Here, we summarize the issues that need to be resolved to allow the application of pose estimation in sports biomechanics.

1) Current of pose estimation models

First, numerous 2D pose estimation models using DL have been developed since the establishment of DeepPose. Particularly in the last decade, the technology in pose estimation has been revolutionized by the excellent work of computer scientists. The accuracy of joint position estimation has also shown great improvement when using the criteria of common evaluation metrics. However, the development trend has shifted from the accurate estimation of joint center positions to multi-person technologies. Accordingly, there is a gap between the technology required by sports biomechanists and the direction of technology development in this field. In sports biomechanics, the focus is often on a single person; therefore, analyzing multiple people is not important. Instead, an important aspect is the precision of common positions of a single person. The pose estimation methods described in this review directly estimate the joint center positions of a person in an image; contrastingly, biomechanical methods indirectly identify the joint centers using anatomical landmarks as cues. Therefore, the joint positions shown by the pose estimation model are in “reasonable” positions based on the teacher data; however, they lack anatomical rigor. This is because the pose estimation model directly estimates the joint center position, unlike a sports biomechanist who marks anatomical landmarks by palpation and calculates the joint center position based on these landmarks[51]. If the pose estimation model could, for example, estimate the ankle joint center position by relying on ankle irregularities, the estimation accuracy would be equivalent to that required by sports biomechanics. There are various techniques for 3D pose estimation; however, in most cases, 2D pose estimation is first performed, followed by 3D estimation. Therefore, the issues experienced in 2D pose estimation exist in 3D pose estimation.

2) Challenges in using existing datasets for sports biomechanics

Those 2D datasets used “in-the-wild” videos and images, with crowd-workers often digitizing the joint center positions; additionally, the development was not aimed at performance or motion analysis. These factors could limit the utility of existing datasets for sports biomechanics. Using social media images, the dataset producer could create a reasonably large amount of teacher data. However, these images did not undergo digitization; therefore, even images with unclear joint positions were considered as digitized to “reasonable” positions. In addition, when digitization was performed using online services, there is a major issue regarding the quality of the crowd-workers. Furthermore, these digitization tasks were not designed for use in sports science. There were many cases where the joint center position was estimated from the clothing or the joint position was clearly off. It is believed that more accurate data sets with clothing will be needed. In order to accurately determine the joint center from above the clothing, there is the way to take measures such as, for example, first applying double-sided tape to the anatomical landmarks and the clothing, and then applying a reflective marker on top of the tape. However, unfortunately, none of the current datasets use this method. Another challenge is that the datasets in 3D were captured using a motion capture system, with less data being included in the dataset than the dataset in 2D created using social media. In addition, since these datasets only include actions that could be performed within the view angle of the motion capture camera, the diversity of the background environment and appearance of the person’s clothes, posture, etc. was relatively low. This may impair the generalization performance of the trained model. To address this problem, various learning methods have been proposed in recent years, including unsupervised learning, semi-supervised learning, self-supervised learning, and weakly supervised learning[55]. As aforementioned, regardless of the model accuracy, the digitization accuracy of the joint center position cannot be sufficiently improved due to the use of existing datasets in the training phase. For the same reason, sports biomechanists need to have detailed knowledge of the datasets used in the markerless motion capture functions that are rapidly becoming popular, such as what sources were used and how they were digitized.

3) Possibility of using existing verification indices in sports biomechanics

This review suggested that existing indices verify the accuracy over the entire video or the average accuracy of the pose estimation for an entire person, moreover, it is difficult to estimate the accuracy of individual joint center position estimation using existing evaluation indices. In sports biomechanics, it is necessary to estimate the exact joint center position for each joint rather than the overall accuracy of the joint position in the image. In fact, several papers have been reported on the accuracy verification of markerless motion capture for use in sports biomechanics[11]. However, most of them used joint position error as an indicator. The validation indicators identified in this study are not of great significance to sports biomechanists, and their use is therefore considered to be low. On the other hand, these metrics still might be useful in areas where more semantic poses can be made without concern for biological or physical constraints, such as human-machine interaction. Nonetheless, when simultaneously evaluating the movements of multiple persons as a whole (e.g., the synchronization degree of group movements in artistic swimming), the existing evaluation indices may still be sufficient.

2 . The current scope of the application of pose estimation models in sports biomechanics

Based on these issues, we consider the scope of application of existing pose estimation models to sports biomechanics. Sports biomechanists can reasonably use these models for recognizing body part location including the termination of movements, which often involves identifying the frame where a certain body part has reached a certain position. For example, Chaudhury et al.[9] used a phasing model in table tennis game analysis to estimate the start and end of a rally scene based on the players' movements. Similarly, pose estimation can be used to record behavior. Sports biomechanists often monitor an athlete's behavior during training to estimate the load on the body, for example, with respect to injury prevention. In this case, the athlete's movements are usually recorded from start to finish, with subsequent visual determination of the duration and frequency of activities. Pose estimation models can perform such a complicated task at a low computational cost. In addition, existing models and evaluation indices can evaluate the overall movements in group activities.

Depth information is theoretically the most difficult to obtain when calculating 3D position coordinates from images. Therefore, the utmost care should be taken when dealing with depth information in 3D pose. Nakano et al.[35] used a sports biomechanical calibration method when calculating the 3D position coordinates of joints from OpenPose[8]. Therefore, it is necessary to use a method similar to theirs when obtaining 3D joint positions by 3D pose estimation. Future development of multi-view images may solve this problem.

3 . What sports biomechanists should do to expand the application of pose estimation models in sports biomechanics

Here, we suggest measures that could be taken by sports biomechanists to expand the utility of pose estimation in sports biomechanics. First, as aforementioned, the current motivation of computer scientists has a different direction than that of movement science, including sports biomechanics, with respect to model development. DeepLabCut[31] is a practical model for movement scientists since it is motivated by the needs of neuroscientists to understand animal behavior. This model allows movement scientists to estimate the behavior of mice and Drosophila with the same accuracy as human digitization. Sports biomechanists should also be proactively involved in the development of technological innovations for improving the accuracy of joint position estimation for a single person. Another major challenge of pose estimation in sports biomechanics is the insufficient learning with existing datasets. Biomechanists should prepare purposeful teacher data. Motion capture systems, which are widely used in sports biomechanics, can overwrite videos with marker information. When using this technology to examine the human body, the video should be recorded; further, information including the positional coordinates of the marker and video should be shared among scientists, which will facilitate the construction of datasets specific to sports biomechanics. However, since the reflection marker is shown in the video that serves as the teacher data, the model may use information regarding the position of the reflection marker to estimate the joint center position[11]. Further research is warranted to examine the effect of marker images on model accuracy when training models using images containing reflective markers. Additionally, there is a need to validate data from systems that can calculate anatomical landmark positions without using markers, including a highly accurate 3D surface analysis system[38], as the gold standard. Regarding model validation, existing evaluation methods are insufficient for determining the exact joint center positions of a single person in sports biomechanics. Finally, when considering the use of pose estimation in sports biomechanics, it is desirable to apply physical quantities (position, velocity, acceleration, etc.) used by sports biomechanists as indicators for evaluation.

Ⅳ. Conclusion

This study sought to clarify the utility of DL-based pose estimation in solving the problems of traditional methods in sports biomechanics. We observed a dramatic improvement in the performance of pose estimation models; however, the direction of development does not meet the needs of movement scientists such as sports biomechanists. Additionally, currently available datasets may contain data with inaccurately digitized joint center positions. The evaluation metrics used for these models were based on the average evaluation of the entire frame or video. Therefore, for sports biomechanists seeking to determine the exact position of individual joint centers, the utility scope of pose estimation may be limited to movement periodization and counting the number of exercises. This is not a systematic study because not all papers on all models were reviewed, but this does not significantly affect the results of this study. This is not only because the accuracy of existing posture estimation models has not been evaluated to the standards required by sports biomechanics, but also because of a lack of appropriate datasets. Computer scientists are exploring new possibilities in pose and, if sports biomechanists wish to utilize this, they should carefully consider the scope of its application. To facilitate the use of pose estimation in sports biomechanics, sports biomechanists should be actively involved in the development of models and datasets; further, the models should be evaluated using physical quantities used by sports biomechanists.

References

1) Amazon. Amazon Mechanical Turk. https://www.mturk.com/ (January 16, 2022)

2) Andriluka M, Iqbal U, Insafutdinov E, Pishchulin L, Milan A, Gall J, Schiele B. PoseTrack: A benchmark for human pose estimation and tracking. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 5167-5176, 2018.

3) Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human Pose estimation: New benchmark and state of the art analysis. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 3686-3693, 2014.

4) Badiola BA, Mendez ZA. A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise. Sensors, 21(18) : 5996, 2021.

5) Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP J Image Video Process, 246309, 2008.

6) Best-in-Class Data Labeling Platform, Playment. https://www.playment.io/ (January 16, 2022)

7) Boser QA, Valevicius AM, Lavoie EB, Chapman CS, Pilarski PM, Hebert JS, Vette AH. Cluster-based upper body marker models for three-dimensional kinematic analysis: Comparison with an anatomical model and reliability analysis. J Biomech, 72(4) : 228-234, 2018.

8) Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D Pose estimation using part affinity fields. Proc IEEE Comput Soc Conf

Comput Vis Pattern Recognit, 1302-1310, 2017.

9) Chaudhury S, Kimura D, Vinayavekhin P, Munawar A, Tachibana R, Ito K, Inaba Y, Matsumoto M, Kidokoro S, Ozaki H. Unsupervised temporal feature aggregation for event detection in unstructured sports videos. 2019 IEEE Int Symp Multimedia, 9-97 , 2019.

10) Chen CH, Ramanan D.3D human Pose estimation = 2D Pose estimation + matching. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 5759-5767, 2017.

11) Colyer SL, Evans M, Cosker DP, Salo AI. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. Sports Med Open, 4, 24, 2018.

12) Fang HS, Xie S, Tai YW, Lu C. RMPE: Regional multi-person pose estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 2353-2362, 2017.

13) Ferrari V, Marín-Jimínez M, Zisserman A. Progressive search space reduction for human Pose estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 1-8, 2008.

14) Güler RA, Neverova N, Kokkinos I. DensePose: Dense human pose estimation in the wild. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 7297-7306, 2018.

15) Hindle BR, Keogh JWL, Lorimer AV. Inertial-Based Human Motion Capture: A Technical Summary of Current Processing Methodologies for Spatiotemporal and Kinematic Measures. Appl Bionics Biomech, 2021:6628320, 2021.

16) Huang YC, Liao IN, Chen CH, Ik TU, Peng WC. TrackNet: A deep learning network for tracking high-speed and tiny objects in sports applications. 16th IEEE Int Conf Adv Video Signal Based Surveill, 1-8, 2019.

17) Insafutdinov E, Andriluka M, Pishchulin L, Tang S, Levinkov E, Andres B, Schiele B. ArtTrack: Articulated multi-person tracking in the wild. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 1293-1301, 2017.

18) Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. Eur Conf Comput Vis, 34-50, 2016.

19) Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell, 36(7) : 1325-1339, 2014.

20) Jhuang H, Gall J, Zuffi S, Schmid C, Black MJ. Towards understanding action recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 3192-3199, 2013.

21) Jiang W, Kolotouros N, Pavlakos G, Zhou X, Daniilidis K. Coherent reconstruction of multiple humans from a single image. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 5578-5587, 2020.

22) Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation. Proc British Mach Vis Conf, 1-11, 2010.

23) Johnson S, Everingham M. Learning effective human pose estimation from inaccurate annotation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 1465-1472, 2011.

24) Kanazawa A, Black MJ, Jacobs DW, Malik J. End-to-End recovery of human shape and pose. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 7122-7131, 2018.

25) Kocabas M, Athanasiou N, Black MJ. Vibe: Video inference for human body pose and shape estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 5253-5263, 2020.

26) Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: A large video database for human motion recognition. Proc Int Conf Comput Vis,

2556-2563, 2011.

27) Leardini A, Cappozzo A, Catani F, Toksvig-Larsen S, Petitto A, Sforza V, Cassanelli G, Giannini S. Validation of a functional method for the estimation of hip joint centre location. J Biomech, 32(1): 99-103, 1999.

28) Li J, Wang C, Zhu H, Mao Y, Fang HS, Lu C. Crowdpose: Efficient crowded scenes Pose estimation and a new benchmark. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 10855-10864, 2019.

29) Liang J, Lin M. Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images. Proc Int Conf Comput Vis, 4351-4362, 2019.

30) Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Zitnick CL. Microsoft COCO: Common objects in context. Eur Conf Comput Vis, 740-755, 2014.

31) Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, Bethge M. DeepLabCut: markerless Pose estimation of user-defined body parts with deep learning. Nat Neurosci, 21(9), 1281-1289, 2018.

32) McCaw ST, DeVita P. Errors in alignment of center of pressure and foot coordinates affect predicted lower extremity torques. J Biomech, 28(8): 985-988, 1995.

33) Mehta D, Sridhar S, Sotnychenko O, Rhodin H, Shafiei M, Seidel HP, Theobalt C. VNect: Real-time 3D human Pose estimation with a single RGB camera. ACM Trans. Graph, 36(4), 1-14, 2017.

34) Monnet T, Desailly E, Begon M, Vallée C, Lacouture P. Comparison of the SCoRE and HA methods for locating in vivo the glenohumeral joint centre. J Biomech, 40(15): 3487-3492, 2007.

35) Nakano N, Sakura T, Ueda K, Omura L, Kimura A, Iino Y, Yoshioka S. Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras. Front Sports Act Living, 2(50), 2020.

36) Newell A, Yang K, Deng J. Stacked hourglass networks for human Pose estimation. Eur Conf Comput Vis, 483-499, 2016.

37) Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler P, Schiele B. DeepCut: Joint subset partition and labeling for multi person Pose estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 4929-4937, 2016.

38) Popat H, Richmond S, Benedikt L, Marshall D, Rosin PL. Quantitative analysis of facial movement-A review of three-dimensional imaging techniques. Comput Med. Imaging Graph, 33(5), 377-383, 2009.

39) Raaj Y, Idrees H, Hidalgo G, Sheikh Y. Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 4615-4623, 2019.

40) Reinschmidt C, Van der Bogert AJ, Nigg BM, Lundberg A, Mur- phy N. Effect of Skin Movement on the Analysis of Skeletal Knee Joint Motion During Running. J Biomech, 30(7): 729-732, 1997.

41) Reno V, Mosca N, Marani R, Nitti M, D'Orazio T, Stella E. Convolutional neural networks based ball detection in tennis games. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 1839-1845, 2018.

42) Sapp B, Taskar B. MODEC: Multimodal decomposable models for human Pose estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 3674-3681, 2013.

43) Seethapathi N, Wang S, Saluja R, Blohm G, Kording KP. Movement science needs different pose tracking algorithms. arxiv preprint, https://arxiv.org/abs/1907.10226, 2019.

44) Sigal L, Balan AO, Black MJ. HumanEva: Synchronized video and motion capture dataset

and baseline algorithm for evaluation of articulated human motion. Int J Comput. Vis, 87, 4-27, 2010.

45) Song L, Yu G, Yuan J, Liu Z. Human pose estimation and its application to action recognition: A survey. J Vis Commun Image Represent, 76, 103055, 2021.

46) Tompson J, Jain A, LeCun Y, Bregler C. Joint training of a convolutional network and a graphical model for human Pose estimation. Adv Neural Inf. Process. Syst, 2, 1799-1807, 2014.

47) Toshev A, Szegedy C. DeepPose: Human Pose estimation via deep neural networks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 1653-1660, 2014.

48) Vafadar S, Gajny L, Boëssé M, Skalli W. Evaluation of CNN-Based Human Pose Estimation for Body Segment Lengths Assessment. ECCOMAS Thematic Conf Comput Vis Med Image Process, 34, 179-187, 2019.

49) Wang J, Tan S, Zhen X, Xu S, Zheng F, He Z, Shao L. Deep 3D human pose estimation: A review. Comput Vis Image Underst, 210, 103225, 2021.

50) Wei SE, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 4724-4732, 2016.

51) Wu G, Siegler S, Allard P, Kirtley C, Leardini A, Rosenbaum D, Whittle M, D'Lima DD, Cristofolini L, Witte H, Schmid O, Stokes I. ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion-part I: ankle, hip, and spine. J Biomech, 35(4) : 543-548, 2002.

52) Xiu Y, Li J, Wang H, Fang Y, Lu C. Pose flow: Efficient online pose tracking. Proc British Mach Vis Conf, 1-12, 2018.

53) Xu J, Yu Z, Ni B, Yang J, Yang X, Zhang W. Deep kinematics analysis for monocular 3D human pose estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 896-905, 2020.

54) Yang Y, Ramanan D. Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach Intell, 35(17) : 2878-2890, 2013.

55) Zhai X, Oliver A, Kolesnikov A, Beyer L. S4L: Self-Supervised Semi-Supervised Learning. Proc Int Conf Comput Vis, 1476-1485, 2019.