Multimedia Tools and Applications (2024) 83:22995–23018
https://doi.org/10.1007/s11042-023-16369-8

Motion-aware and data-independent model based multi-view 3D pose refinement for volleyball spike analysis

Yanchao Liu[1] · Xina Cheng[2] · Takeshi Ikenaga[1]

Received: 3 October 2022 / Revised: 25 May 2023 / Accepted: 17 July 2023 /
Published online: 10 August 2023

Abstract

In the volleyball game, estimating the 3D pose of the spiker is very valuable for training and analysis, because the spiker's technique level determines the scoring or not of a round. The development of computer vision provides the possibility for the acquisition of the 3D pose. Most conventional pose estimation works are data-dependent methods, which mainly focus on reaching a high level on the dataset with the controllable scene, but fail to get good results in the wild real volleyball competition scene because of the lack of large labelled data, abnormal pose, occlusion and overlap. To refine the inaccurate estimated pose, this paper proposes a motion-aware and data-independent method based on a calibrated multi-camera system for a real volleyball competition scene. The proposed methods consist of three key components: 1) By utilizing the relationship of multi-views, an irrelevant projection based potential joint restore approach is proposed, which refines the wrong pose of one view with the other three views projected information to reduce the influence of occlusion and overlap. 2) Instead of training with a large amount labelled data, the proposed motion-aware method utilizes the similarity of specific motion in sports to achieve construct a spike model. Based on the spike model, joint and trajectory matching is proposed for coarse refinement. 3) To finely refine, a point distribution based posterior decision network is proposed. While expanding the receptive field, the pose estimation task is decomposed into a classification decision problem, which greatly avoids the dependence on a large amount of labelled data. The experimental dataset videos with four synchronous camera views are from a real game, the Game of 2014 Japan Inter High School of Men Volleyball. The experiment result achieves 76.25%, 81.89%, and 86.13% success rate at the 30mm, 50mm, and 70mm error range, respectively. Since the proposed refinement framework is based on a real volleyball competition, it is expected to be applied in the volleyball analysis.

Keywords Data independence · Motion-aware model · 3D human pose refinement · Sports analysis in volleyball

✉ Yanchao Liu
liuyanchao@fuji.waseda.jp

Extended author information available on the last page of the article

1 Introduction

With the development of computer vision, 3D human pose is increasingly used in sports analysis, and human-computer interactions, and animation. Unlike 2D poses limited to the image or video domain, 3D poses contain multi-level information in the real world. In the volleyball field, at the spatial level, the 3D pose shows the position and velocity of players, which are used to analyze the team formation and the players' status. At the temporal level, a series of the player's 3D poses reflect which motion the player is doing, such as serving, spiking, or receiving. This information helps the audience understand the motion in a volleyball game for TV broadcasting. Furthermore, at the kinematics level, using the generalization feature of 3D poses, the coach or even the computer [8] is able to guide the player's motions by comparing the player's 3D pose with a series of standard 3D poses to improve the training efficiency.

In a whole volleyball game, Table 1 shows action terms in the volleyball game. In these actions, the efficacy of the spike is the most important action [28], because the spike directly determines the scoring or not of a round. The definition of the spike is the action of attacking the ball over the net with force and intent to score a point. In order to better analyze the motion and improve the technical level of the spiker, the 3D pose of the spiker is essential. Therefore, this paper is aiming to refine the 3D pose of the spiker.

With the rapid development of deep learning, such solutions have been shown in various tasks including image recognition [15, 31, 45], action recognition [6, 21, 42], and image steganography [19, 26, 27]. Significant progress and remarkable performance have already been made by employing deep learning techniques in 3D pose estimation tasks. The goal of 3D human pose estimation is to localize joint keypoints of a human body in real space. Dong et al. [9], Pavlakos et al. [29], Iskakov et al. [18], Kocabas et al. [20], and Qiu et al. [30] achieve the high performance on dataset with controllable scene. However, conventional pose estimators fail to get good results on the wild and real volleyball game scene, which has the following challenges: (1) The pose of the spiker is very different from the common pose in daily life scenes. The abnormal pose mainly causes miss joint point error, which represents a large displacement from the groundtruth joint position; (2) There are not only 12 players on a large volleyball court but also referees and staff. The occlusion and overlap mainly lead to the inversion or swap joint points error. Inversion error occurs when a pose estimation model is confused between semantically similar parts that belong to the same instance, such as left

Table 1 Terms in the volleyball game. Note that this paper is aiming to refine the 3D pose of **spike** action

Terms	Definition
Serve	The action of starting a rally by sending the ball over the net to the opposing team.
Pass	The action of receiving the ball from a serve or attack, usually with the forearms.
Set	The action of placing the ball in the air for a teammate to hit over the net.
Spike	The action of attacking the ball over the net with force and intent to score a point.
Block	The action of stopping an opponent's attack at the net by jumping and reaching above the net with the hands.
Dig	The action of diving or reaching to save a ball that has been hit by the opposing team, usually with the forearms.

or right elbow exchange; Swap error represents a confusion between the same or similar parts which belong to different persons; (3) The application scene is for real competitions, similar to the healthcare field [14], there is no sufficient labelled data for training or fine-tuning. These challenges cause estimated errors from conventional pose estimation works (Fig. 1).

In order to overcome these errors caused by the challenges mentioned above for accurate 3D pose, utilizing multiple views information at the spatial level and the multiple frames continuity at the temporal level is the key. In recent years, an increasing number of efforts in the community are focused on pose refinement:

At the spatial level, Moon et al. [25] combine the empirical pose error distributions to propose a model-agnostic pose refinement network for common pose estimation work. Fieraru

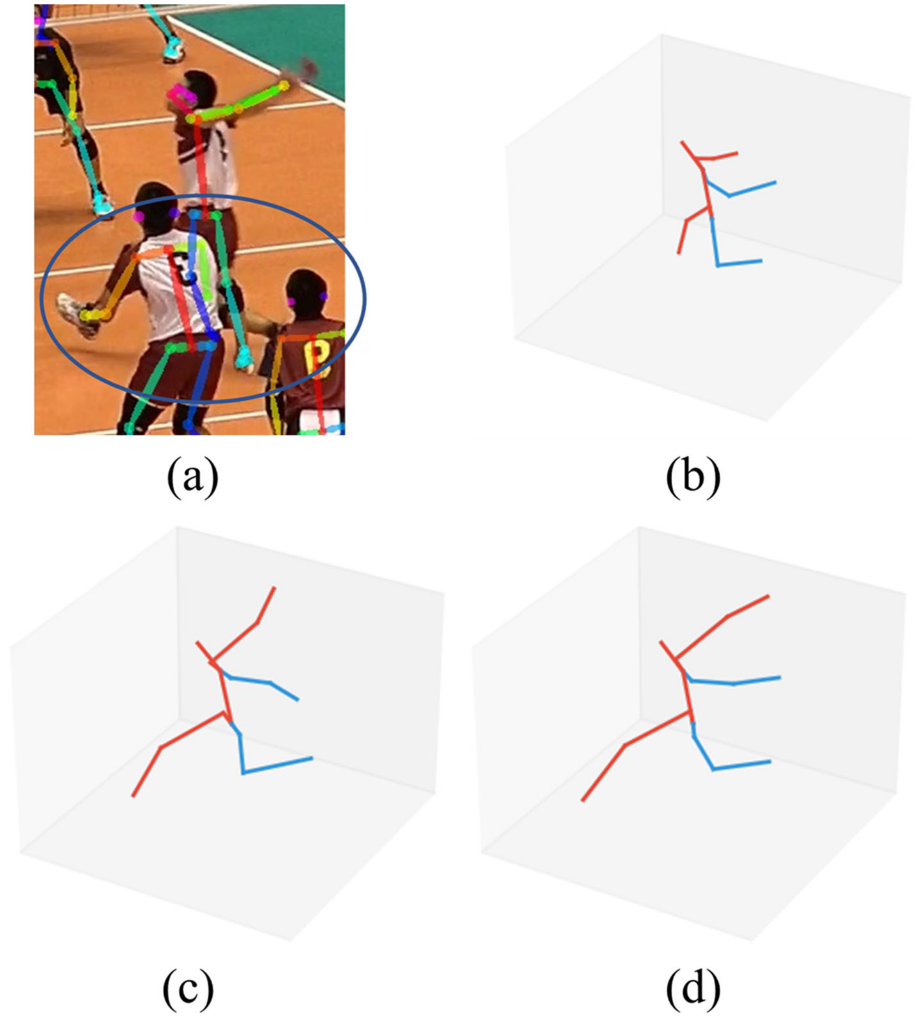


Fig. 1 **a** shows the 2D pose estimated by OpenPose [4] from cropped frame (horizontal flip for better vision) of one view, the ellipse area shows that the knee and ankle joints are not successfully detected. **b** shows the reconstructed 3D pose with 2D estimated results. **c** shows our refined 3D pose. **d** shows the groundtruth 3D pose

et al. [11] directly generate the refined body pose from the initial pose prediction in one forward pass by exploiting the dependencies between the input and output spaces. D'Eusanio et al. [7] refine the 3D human pose with a depth map and a coarse 2D pose. The above methods explore effective and general pose refinement at the spatial level. However, these methods are based on a single image, and cannot benefit from the effectiveness of multiple views. Moreover, for the continuous spiking motion of volleyball, in addition to the spatial intra-frame information at the image level, the temporal continuity is very meaningful to be considered.

At the temporal level, considering the temporal continuity of video, the pose of the human body does not suddenly change or disappear, there are rich correlations between the multi-frames, and it is meaningful to refine the human pose by considering the prior temporal information. A non-linear pose manifold [24, 41] model is constructed by using temporal prior. Wang et al. [41] propose a method that learns a number of bases to obtain tight approximations of the low-dimensional pose manifold, which is able to avoid generating illegitimate poses. With the pose manifold model, Mei et al. [24] refine noisy 3D human pose sequences by jointly projecting them onto a non-linear pose manifold. These works explore good generalization results in the field of video-based 3D human pose refinement. However, these two methods directly refine the 3D pose from input 3D pose sequences and do not utilize multi-view information. Moreover, these data-dependent methods mentioned above require a large amount of labelled data, which is difficult to apply to the real wild volleyball competition scene. Tian et al. [38] combine temporal prior information and joints heatmap to refine the 2D skater's pose, utilizing multi-constrains by skating knowledge to refine the skater's pose in 3D space. But this work only refines the 2D pose on each view one by one and also does not consider the relationship of multiple views. To overcome the limitations of the existing method, simultaneously considering multi-view relationships, prior temporal information in a unified framework is highly desired.

Multi-view Relationships The proposed framework includes an irrelevant projection based potential joint restore method. The main idea is that utilize the projection relationship of the multi-view camera system, aiming to restore joint points with swap or inversion errors mainly caused by occlusion or overlap. The logic behind this method is that the joint point with swap or inversion error is not completely undetected or wrong, but is just given the wrong label, such as a different person ID or a different joint point number. In a single-view system, it may be difficult to refine, but in a multi-view system, these errors in the one view are able to be refined by projecting the reconstructed results from other pairwise views.

Prior Temporal Information To overcome the problem of insufficient labelled data, a motion-aware pose model is constructed by combining the spike motion and temporal prior information with limited labelled sequences. The intuition behind this proposal is that the same type of motion in different sequences is highly similar in sports. For instance, such as jumping in figure skating or spiking in volleyball, although they are abnormal and difficult to estimate correctly, the similarity is regarded as temporal prior information to guide the refinement. This intuition is the key to getting rid of data dependence. Based on the spike model, with joint and trajectory matching, the missing error mainly caused by abnormal pose is refined to the coarse-refined point. To obtain more accurate points, a point distribution based posterior decision network is proposed for fine refinement. Specifically, several 3D points are used to generate a distribution around the coarse-refined points, aiming to generate a larger receptive field. A posterior decision multi-network is proposed to filter the distribution to get the accurate refined points. Therefore, the complex estimation problem is transformed into a pure classification decision problem, which only requires a limited number of labelled data for training to overcome data dependence.

In particular, the contributions of this paper are briefly summarized as follows:

1. We propose an innovative projection-based potential joint restore method for multi-view camera systems. This method leverages the benefits of multi-view reconstruction and projection to enhance the accuracy of joint point inversion and swap errors, which are primarily caused by occlusion or overlap in individual views.
2. We propose a novel motion-aware spike model based matching method to effectively correct the miss errors caused by abnormal poses in real-world volleyball game scenes. This method utilizes prior motion and temporal information to refine the 3D poses. In the context of volleyball, different sequences exhibit highly similar motion patterns. This motivates us to leverage the temporal prior of poses, requiring only a small amount of labeled data in advance to reduce data dependency.
3. To further enhance the refinement performance, we propose a point distribution based posterior decision network. The novelty is that the proposed method decomposes the pose estimation task into a pure classification decision problem. By employing this method, we significantly reduce the reliance on a substantial amount of labeled data, thus achieving fine refinement with improved accuracy.
4. The experiment results show that the proposed method achieves a better result than the conventional pose estimation works on a wild real volleyball game dataset, reflecting that the proposed method is able to be applied to actual volleyball matches.

The rest of this paper is arranged as follows. The related work is introduced in Section 2. Section 3 shows the framework and the detail of the proposed methods. The experiment result is described in Section 4. Finally, Section 5 is the conclusion.

2 Related work

2.1 Sports analysis with human pose

The player pose plays a very important role in the sports analysis field. By analyzing the pose of the athlete, a lot of data that is very helpful for sports analysis is able to be obtained. Cheng et al. [5] propose a spike height analysis system, which combines the pose and heatmap of the player to obtain the spike height of 3D space in the real world for volleyball game analysis. Askari et al. [2] propose a Recurrent Neural Network (RNN) based method, equipped with image features and players' poses as input to recognize the interaction for ice hockey analysis. Zhu et al. [47] propose FenceNet with 2D pose input to automate the classification of fine-grained footwork techniques in fencing. Analyzing the pose of the athlete is also used as a training aid. Wang et al. [40] propose a spatial-temporal relation module, considering the spatial relation of different keypoints among each time frame and temporal relation of specific keypoints among time dimensions simultaneously, aiming to achieve abnormal detection and exemplar-based visual suggestions for a better user training experience. Zou et al. [48] present a fitness training system with pose estimation, which not only shows fitness training courses but also provides motion correction. Guo et al. [13] design a usable visual analytic prototype with pose estimation for cheerleading and dance training. Liu et al. [22] design a golf swing evaluation system, which uses the similarity between the player pose and reference pose to evaluate swing quality and report a score ranging from 0 to 10. The above work reflects that accurate and reliable athlete pose is extremely important for sports analysis. Therefore, this paper focuses on refining the performance of general pose estimation works for spiking motion in volleyball games.

2.2 Multi-view 3D pose estimation

With the development of deep learning, The requirements for the accuracy of pose estimation are also increasing. For the calibrated multi-view camera system, most approaches estimate 2D poses with monocular pose estimators [1, 4, 10] from each view separately to reconstruct or recover the 3D poses. As for 2D pose estimation, Cao et al. [4] propose OpenPose, which is a bottom-top estimator, through heatmaps and part affinity fields, using convolutional pose machines to predict joint coordinates and connection of multi-person in one image. Fang et al. [10] propose a top-down regional multi-person pose estimation method, which contains a symmetric spatial transformer network, parametric pose non-maximum-suppression, and pose-guided proposals generator. Artacho and Savakis [1] propose OmniPose, which incorporates contextual information across scales and joint localization with Gaussian heatmap modulation at the multi-scale feature extractor to estimate human pose. As for 3D recovering or reconstruction, Dong et al. [9] propose using the combination of geometric, appearance, and cycle-consistency constraints to design a matching algorithm to reconstruct 3D poses with clustering 2D poses. Pavlakos et al. [29] reconstruct the 3D poses from multi-view 2D poses by gathering 3D annotations with a ConvNet for 2D pose estimation, and recordings from a multi-view setup. Iskakov et al. [18] propose a learnable method that combines a basic differentiable algebraic triangulation and volumetric aggregation from intermediate 2D backbone feature maps. Kocabas et al. [20] utilize epipolar geometry to reconstruct 3D poses and camera geometry from multi-view 2D poses to train a 3D pose estimator. Qiu et al. [30] propose a CNN based multi-view feature fusion approach to improve the 2D pose estimation accuracy, and recover 3D poses from multi-view 2D poses by incorporating multi-view geometric priors in the model. Specifically, these 2D poses are estimated from multi-view images, captured by cameras from different angles at the same time. And then they present a recursive Pictorial Structure Model to recover the 3D pose from the multi-view 2D poses.

The existing pose estimation approaches mainly depend on the massive labelled data to focus on achieving great performance on open-source datasets with controllable scene, but for the wild real volleyball scene, with the limitation of large amount of labelled data, it is worth to further exploring a data-independent refinement method.

2.3 Pose refinement

There still exist a lot of difficult cases such as the volleyball scene where even the state-of-the-art method cannot estimate all joints without any error. Based on the conventional pose estimation framework, many works are focusing on refining the estimated pose to increase detection accuracy. Moon et al. [25] propose a model-agnostic pose refinement method called PoseFix, which uses error statistics as prior information to generate synthetic poses to train the model. Fieraru et al. [11] propose a solution by directly generating the refined body pose from the initial pose prediction in one forward pass, exploiting the dependencies between the input and output spaces. For the video input, the multi-frames temporal prior information is widely utilized. A non-linear pose manifold [24, 41] model is constructed to refine the 3D pose. Zhou et al. [46] propose a temporal keypoint matching and refinement network by matching keypoints in across frames and matching pose in adjacent frames. Véges and

Lőrincz [39] propose an energy minimization approach for the smooth, valid trajectories in time. Zeng et al. [43] propose a plug-and-play refinement network, which suppresses the influence of long-term jitters from the initial pose result. For the multi-view video input, Tian et al. [38] propose a multi-task architecture based on a calibrated multi-camera system to facilitate jointly 3D jump pose of figure skater, which uses temporal heatmap on 2D by temporal prior information and multi-constrains on 3D by skating prior knowledge to refine the inaccurate pose results. Considering the relationship of multi-view cameras, Bridgeman et al. [3] propose a method to associating poses between multi-view works by seeking the best correspondences first in a greedy fashion, while reasoning about the cyclic nature of correspondences to constrain the search. Based on the knowledge of volleyball, this paper constructs a spike model to refine the wrong joint points by using the similarity of spike action. Existing pose refinement works are mainly data-dependent and utilize generalized prior information and made indelible achievement. But for spike motions in volleyball game, lacking labelled data makes training difficult, and generalized priors do not fully describe the motion because of their abnormality. We observed that the same motions in different sequences are abnormal but have a high similarity, this paper proposes to build a spiking motion-aware and data-independent model as a motion prior to guide pose refinement.

3 Proposal

3.1 Preprocessing

The input is a set of synchronous 4 views spike videos $\mathbf{V}$ with 4K resolution and 30 Frames Per Second, and a series of estimated 2D poses $\mathbf{x}$ of the spiker. The 2D poses are obtained by the conventional 2D pose estimators. A well-known high-accuracy multi-person 2D pose estimation work called OpenPose [4] is implemented as the baseline to obtain the 2D pose of the spiker. Specifically, For each video in $\mathbf{V}$, firstly, we applied cropping to the video based on the spiker's position in order to eliminate extraneous information. Then we input cropped videos to the pre-trained OpenPose demonstration. Given that the poses generated by OpenPose are anonymous, meaning that poses associated with the same person ID in different frames may not necessarily pertain to the same individual, it becomes essential to filter the initial poses. We utilize a simple and efficient method to filter the 2D pose of the spiker from the initial poses. We annotate the pose of the spiker in the first frame. In the subsequent frames, we establish the association of the pose by calculating the distance between the positions of joint points across consecutive frames. Among the all initial poses, we identify the pose with the minimum distance from the previous frame as the current frame's spiker pose. To ensure reliability, we use joint points that are not prone to errors for calculating the distance: nose, neck, shoulders, and hips joints.

In volleyball analysis, we pay more attention to the movements of the limbs rather than facial organs or toes. Therefore, we remove the joint points of the eye, ear, toe, and heel from BODY25 pose model[1] generated by OpenPose. In this paper, the joint points from number 0 to number 14 of the BODY25 model are considered.

[1] The joint of the BODY25 model uses 25 joint points to represent the human body, and the joint numbers are shown as follows: 0 Nose, 1 Neck, 2 RShoulder, 3 RElbow, 4 RWrist, 5 LShoulder, 6 LElbow, 7 LWrist, 8 MidHip, 9 RHip, 10 RKnee, 11 RAnkle, 12 LHip, 13 LKnee, 14 LAnkle, 15 REye, 16 LEye, 17 REar, 18 LEar, 19 LBigToe, 20 LSmallToe, 21 LHeel, 22 RBigToe, 23 RSmallToe, 24 RHeel. L means left, and R means right.

Finally, the input 2D pose of the spiker is presented as $\mathbf{x} = \left\{\mathbf{x}_c \in \mathbb{R}^{F\times N\times 2}\right\}_{c=1}^{4}$, which c is the index of camera view, F is total frames of input video, $N = 15$ is 15 joint points in the BODY25 model, 2 is the 2D coordinate of the joint point in each frame.

3.2 Framework

Figure 2 shows the overall framework of our 3D pose refinement system. The goal of this work is to obtain a serial of refined 3D pose $\mathbf{X^r}$, i.e.

$$\mathbf{X^r} \in \mathbb{R}^{F\times N\times 3} = \Gamma(\mathbf{V}; \mathbf{x}), \tag{1}$$

where Γ is an abstraction that represents the overall framework, the meaning of F and N are the same as the 2D pose, 3 is the 3D coordinate of the joint point in the real world, $\mathbf{x}$ is the 2D pose obtained by the conventional 2D pose estimator and $\mathbf{V}$ is given multi-view spike videos.

The proposed framework designs a three-stage method with 2D-to-3D and coarse-to-fine strategy to refine the pose of the spiker. In stage one, as we discussed, the swap or inversion error is not completely undetected or wrong, but is just given the wrong label. As for the multi-view system, the 2D swap and inversion errors for a single view are refined by exploiting information from the other three irrelevant views. These three irrelevant views are combined by pairs for triangulation reconstruction [16], and dotted lines of the same color protruding from the three irrelevant views indicate pairings in Fig. 2. The obtained three 3D points are reprojected to another view, which are the three white points in Fig. 2. These three projected points are supplementary information to find out whether there are potential joint points that can be recovered. Finally, the RWrist point of the other player is restored as RAnkle of the spiker. In stage two, a series of multi-view 2D poses are reconstructed into 3D poses. Besides, with the motion-aware and temporal prior information, a spike model is constructed with few labelled sequences. By comparing the reconstructed 3D poses and 3D poses in the spike model on joint and trajectory level, a matching relationship is able to be calculated to transform the 3D poses in the model to the reconstruction result, the miss joint points with large deviations are refined to small deviations. After, the coarse-refined 3D poses are input to stage three to do the finely refinement. In stage three, for each joint in the coarse-refined 3D pose, a point set is distributed around each joint point in the 3D space. A decision network estimates the confidence score for each distributed point to find the best one as the final 3D joint point. After the operation frame by frame, a series of finely refined 3D poses are output finally. Each proposed method is presented in the following subsections.

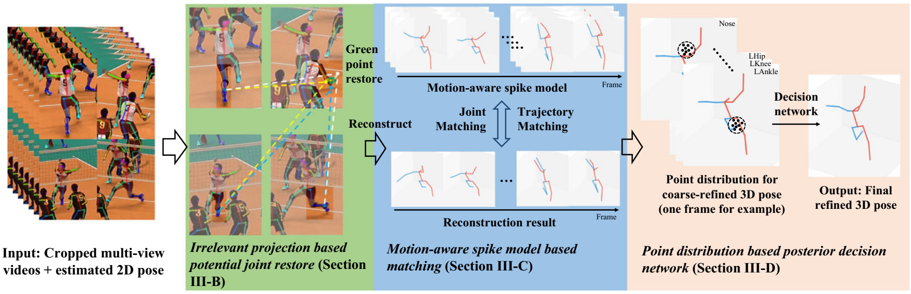

Input: Cropped multi-view videos + estimated 2D pose
Green point restore
Reconstruct
Irrelevant projection based potential joint restore (Section III-B)
Motion-aware spike model
Frame
Joint Matching
Trajectory Matching
Reconstruction result
Frame
Motion-aware spike model based matching (Section III-C)
Nose
LHip
LKnee
LAnkle
Decision network
Point distribution for coarse-refined 3D pose (one frame for example)
Output: Final refined 3D pose
Point distribution based posterior decision network (Section III-D)

Fig. 2 Framework of the proposed 3D pose refinement method, in which a data-independent, 2D-to-3D, and coarse-to-fine architecture is proposed with combining the multi-view relationship, motion and temporal prior

3.3 Irrelevant projection based potential joint restore

The accuracy of the 3D pose estimation is highly dependent on the corresponding 2D pose estimation. Hence, refining the 2D pose is of utmost importance prior to refining the 3D pose. Refining the 2D pose based on a single image is a challenging task. However, in the case of multi-view systems, the availability of additional information from other views proves advantageous. Consequently, we propose an innovative projection-based potential joint restore method for the multi-view camera system, refining the 2D pose for each view. Benefiting from other views, it becomes feasible to refine mislabeled errors, such as swap and inversion, which constitute nearly half of all errors [25].

To end this, with a four-view system, we consider that for the joint point with the same number, use three views that are irrelevant to the current view. Then we reconstruct pairwise these three views into 3D space and then reproject them back to the current view. And we calculate the Euclidean distance between the estimated joint point and reprojected points. Then, a threshold τ_1 is used to determine whether there is an error joint point in the current view. If the error happens and the estimated result of the other three irrelevant views is good, all detected joint points of all players are searched. Through evaluating the Euclidean distance between each joint point and reprojected points one by one, comparing with a threshold, the point with the smallest distance and less than the threshold τ_2 is restored. Formally, considering the frame f in camera view c of $\mathbf{V}$, the irrelevant reconstructed 3D points $\mathbb{R}^I$ are defined as

$$\mathbb{R}_c^I[f] = \left\{ \mathbf{X}_{(i,j)}^I[f] \in \mathbb{R}^{N\times 3} \right\}_{i\neq j\neq c}, \tag{2}$$

where the i, j, c represent the index of camera view from 1 to 4 in four-view system, I means irrelevance. For a camera view c, there are three irrelevant camera pairs represented as (i, j). Then, the reprojection $\mathbb{P}^I$ is defined as

$$\mathbb{P}_c^I[f] = \left\{ \mathbf{x}_{(i,j)}^I[f] \in \mathbb{R}^{N\times 2} \right\}_{i\neq j\neq c}. \tag{3}$$

For each joint, the Euclidean distance between the estimated joint point and irrelevant reprojection points is defined as

$$\mathbf{D}_c^{EI}[f,n] = \left\{ \left\| \mathbf{x}_c[f,n],\ \mathbf{x}_{(i,j)}^I[f,n] \right\|_2 \right\}_{i\neq j\neq c}, \tag{4}$$

where n is the joint number from 0 to 14, and there are three element in $\mathbf{D}_c^{EI}[f,n]$ for four-view system. A threshold $\tau_1 = 40$ is used to evaluate the each distance in $\mathbf{D}_c^{EI}[f,n]$, a score is given to each view c for each joint n, which is shown as:

$$Score_c[f,n] = \sum \mathbf{Th}(\mathbf{D}_c^{EI}[f,n], \tau_1), \tag{5}$$

where the $\mathbf{Th}$ is the threshold function, if the distance is equal or less than the τ_1, the output is 1, else 0, and three elements in $\mathbf{D}_c^{EI}[f,n]$ are compared with τ_1 one by one, the output is summed as $Score_c[f,n]$. After combining the scores of the four views, there are the following three cases. 1) The first case is that each view's score is equal to 3, which shows the three irrelevant projected points are close to the estimated point. It means that the estimated points of the four views have the correct correspondence, so the joint point n of frame f is marked as correct in all four views. 2) The second case is that each view's score is equal to 0, which shows all the irrelevant projected points are not close to the estimated point. It means that at least two views have wrong estimated points, but cannot distinguish which view has a wrong point, so the joint point n of frame f is marked as to-refine in all four views. 3) The third case is that the score of three views is equal to 1, and the score of another is equal to 0, meaning the three views with score 1 have the correct correspondence. Because the joint point in the one view with score 0 is wrong, the distance between the estimated point and three irrelevant projected points is larger than the threshold τ_1. Therefore, the joint point n of frame f in the one view with score 0 is marked as the to-restore joint (see Fig. 3).

For these to-restore joint points, a global searching method is used to find the best joint to restore. For view c, frame f, joint n, defining the overall joint points in view c as potential joint points $\mathbb{P}_c^P[f,n] = \left\{\mathbf{x}_c^P[f,n] \in \mathbb{R}^{N^P \times 2}\right\}$, where N^P is the total number of potential joint points. Then evaluate the Euclidean distance between each potential joint point and three irrelevant projected points, the point index n^P with minimum distance $\mathbf{D}_c^{PI}[f,n,n^P]$ is defined as

$$\mathbf{D}_c^{PI}[f,n,n^P] = \underset{n^P \in [0,N^P) \in \mathbb{N}}{\arg\min} \left[\frac{1}{C_3^2} \sum_{i \neq j \neq c}^{4} \left\| \mathbf{x}_c^P[f,n,n^P],\ \mathbf{x}_{(i,j)}^I[f,n] \right\|_2 \right], \tag{6}$$

where C is the combination function. And the threshold $\tau_2 = 25$ is used to evaluate the $\mathbf{D}_c^{PI}[f,n,n^P]$ is restored or not. If the distance is less than or equal to the τ_2, the potential joint point n^P is restored and marked as corrected.

$$\mathbf{x}_c[f,n] \leftarrow \mathbf{x}_c^P[f,n,n^P], \tag{7}$$

else the joint point $\mathbf{x}_c[f,n]$ is marked as to-refine, and refined by the subsequent steps, which are presented in the following subsections.

3.4 Motion-aware spike model based matching

In addition to leveraging the multi-view relationship at the 2D level, our work takes into account the specific motion patterns observed in sports as prior knowledge. Limited by the availability of labeled data, data-driven methods are not suitable in this scenario. Unlike conventional data-driven approaches, we recognize that the same motion in different sequences may appear abnormal but exhibits similarity. This similarity can be utilized as meaningful prior temporal information. Motivated by this observation, we propose a novel motion-aware spike model based matching method. This method is based on a contrastive idea, using a data-independent spiking model and test poses as input at the same time, effectively refining the wrong test poses.

In this regard, as shown in Fig. 4, our approach utilizes only a few non-testing spike sequences with labeled 3D poses as reference sequences. These sequences are aligned to generate a spike pose model. This model captures the relative position and distribution at the spatial level, as well as the movement trend and trajectory at the temporal level. Formally, the spike pose model is defined as

$$\mathbb{M}^S = \left\{\hat{\mathbf{X}}^S \in \mathbb{R}^{R \times F_r \times N \times 3}\right\}, \tag{8}$$

where $\hat{\mathbf{X}}$ is the 3D groundtruth of the reference sequence, S means spike, R is the total number of reference sequences, and F_r is the total frames of each reference sequence. The spike model is a series of labelled 3D poses.

The four-views 2D test results get from Section 3.3 are reconstructed to the 3D by triangulation method [16]. The reconstructed 3D poses before refinement are defined as $\mathbf{X} \in \mathbb{R}^{F \times N \times 3}$. Considering both spatial and temporal levels, a Joint and Trajectory Matching

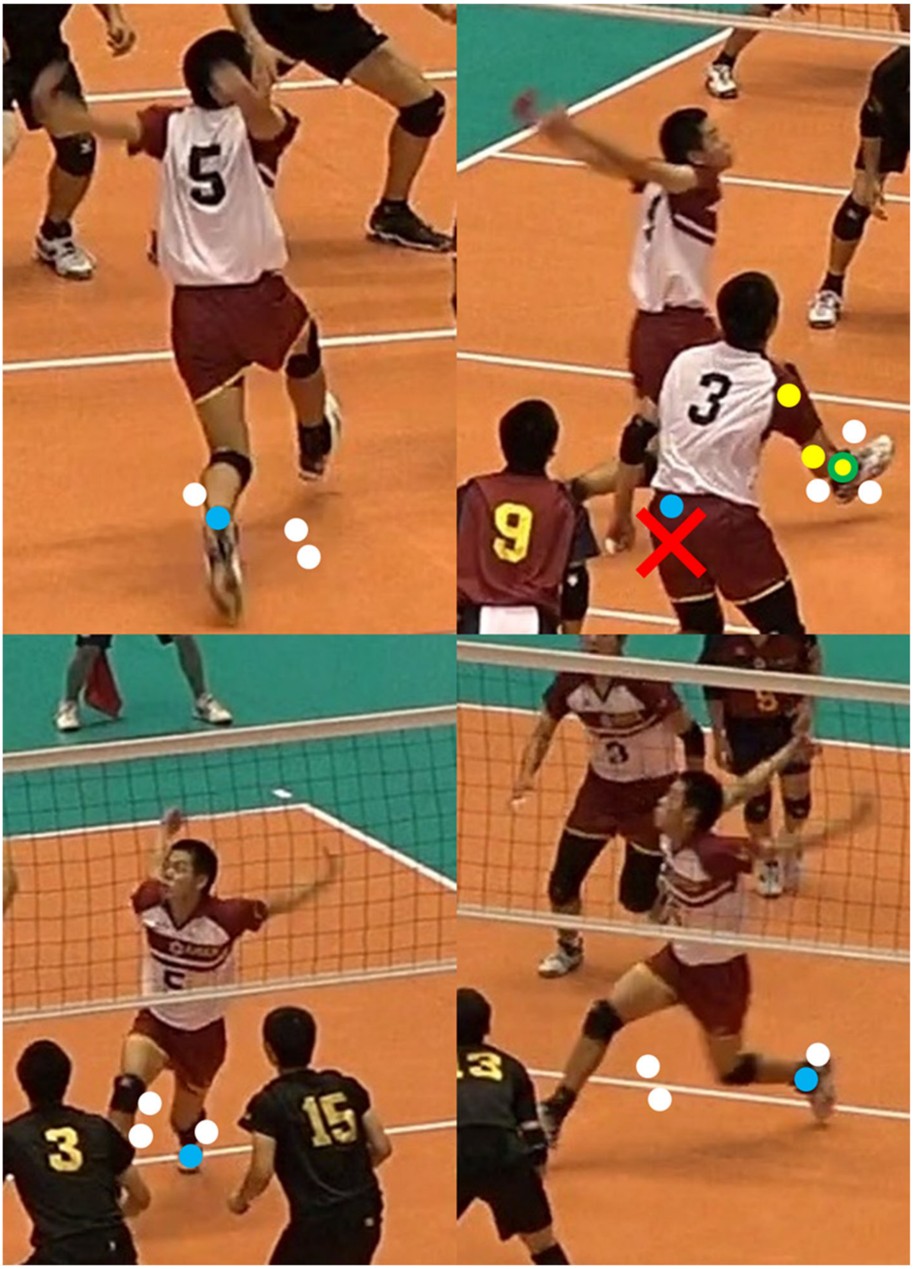


Fig. 3 The schematic diagram of irrelevant projection based potential joint restore: The blue points are estimated LAnkle joint points, the white points are irrelevant projected points, the yellow points are the potential joint points, and the yellow point with green contour is the restored joint point

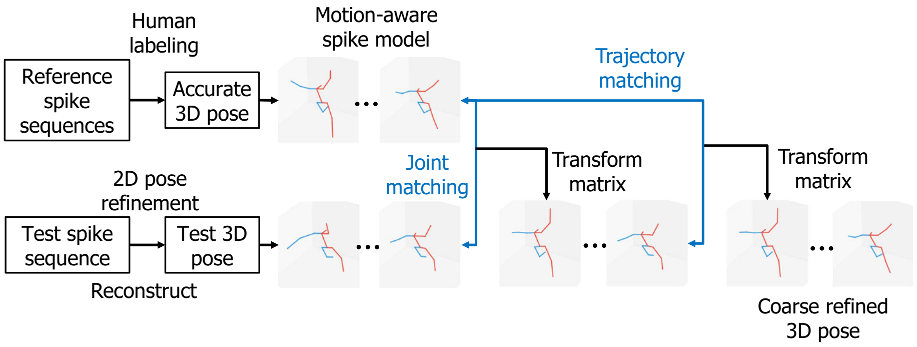


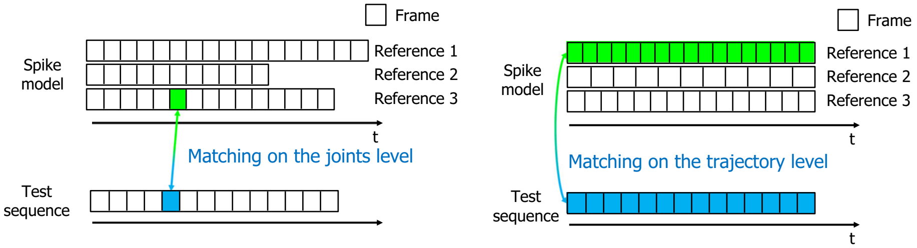
Fig. 4 The concept and schematic diagram of motion-aware spike model based matching

method based on the generic one-sided Procrustes Analysis (JTMPA) [12] is proposed for 3D coarse refinement. Mathematically, the purpose is to find an optimal transformation that makes two matrices as close as possible to each other. Joint and trajectory matching are presented in the following subsections in detail.

3.4.1 Joint matching

Considering the spatial similarity of human poses in the same motion between different sequences, such as the relative position and distribution of joint points, the spike model provides a library of similar poses, and the JTMPA method is proposed to search for the best matching and the transformation matrix. As Fig. 5(a) shows, for the test 3D pose of frame f, to ensure the accurate matching without influence of wrong joint points, choosing the N^J 3D coordinate with corrected mark get from Section 3.3 to concatenate a test joint matrix $\mathbf{X}^J[f] \in \mathbb{R}^{N^J \times 3}$. The Procrustes method requires the shape of the reference joint matrix to be the same as the test joint matrix, choosing the same N^J 3D coordinate in spike model $\mathbb{M}^S$, the spike model with chosen joint points is defined as $\hat{\mathbf{X}}^{SJ} \in \mathbb{R}^{R \times F_r \times N^J \times 3}$. Then a optimized problem is defined that for each 3D pose in the spike model, the best matched 3D pose index r^J of reference sequences, frame index f_r^J and the transformation matrix $\mathbb{A}^J \in \mathbb{R}^{3 \times 3}$ is searched by

$$\mathop{\arg\min}_{r^J,\ f_r^J,\ \mathbb{A}^J, \mathbb{B}^J} \left\| \hat{\mathbf{X}}^{SJ}[r^J,\ f_r^J] \cdot \mathbb{A}^J + \mathbb{B}^J,\ \mathbf{X}^J[f] \right\|_F^2, \tag{9}$$



(a) Joint matching (b) Trajectory matching

Fig. 5 Joint matching and trajectory matching between spike model and test sequence

where, the $\| \ \|_F$ denotes the Frobenius norm, and the matrix $\mathbb{B}^J \in \mathbb{R}^{1\times 3}$ is a translation matrix, which ensures that the reference pose and the test pose are matched in the unified 3D coordinate system. After all, the left $(15 - N^J)$ wrong joint points are refined by the following equation:

$$(\mathbf{X}[f] \setminus \mathbf{X}^J[f]) \leftarrow (\hat{\mathbf{X}}^S[r^J, f_r^J] \setminus \hat{\mathbf{X}}^{SJ}[r^J, f_r^J]) \cdot \mathbb{A}^J + \mathbb{B}^J, \tag{10}$$

where, \ means complement set. After joint matching frame by frame, the refined 3D joint points are reprojected to the view with a to-refined mark, and the mark is updated with Section 3.3.

3.4.2 Trajectory matching

Similar to joint matching, consider the temporal similarity of human pose in the same motion between different sequences. For each joint number n in the test 3D pose, to ensure accurate matching without the influence of wrong joint points, choosing the F^T 3D coordinate with updated corrected mark get from Section 3.3 to concatenate a test trajectory matrix of joint n, which is defined as $\mathbf{X}^T[n] \in \mathbb{R}^{F^T\times 3}$. To ensure the same shape with test trajectory of joint n, the frame length F_r of the reference sequences in the spike model are interpolated to frame length F, and then choosing the same F^T 3D coordinate in spike model $\mathbb{M}^S$, the spike model after choosing is defined as $\hat{\mathbf{X}}^{ST}[n] \in \mathbb{R}^{R\times F^T\times 3}$. Then a optimized problem is defined that for each trajectory of joint n in the spike model, the index r^T of reference sequences, the transformation matrix $\mathbb{A}^T \in \mathbb{R}^{3\times 3}$, and the translation matrix $\mathbb{B}^T$ is searched by

$$\underset{r^T,\ \mathbb{A}^T, \mathbb{B}^T}{\arg\min} \left\| \hat{\mathbf{X}}^{ST}[r^T, n] \cdot \mathbb{A}^T + \mathbb{B}^T,\ \mathbf{X}^T[n] \right\|_F^2, \tag{11}$$

After matching, the left $(F - F^T)$ wrong joint points are refined by the following equation:

$$(\mathbf{X}[n] \setminus \mathbf{X}^T[n]) \leftarrow (\hat{\mathbf{X}}^S[r^T, n] \setminus \hat{\mathbf{X}}^{ST}[r^T, n]) \cdot \mathbb{A}^T + \mathbb{B}^T. \tag{12}$$

After trajectory matching joint by joint, the refined 3D joint points are reprojected to the view, which has the to-refined mark, and the mark is updated with Section 3.3 again.

The overall JTMPA method is shown in Algorithm 1. Since the JTMPA method is based on the motion-aware and temporal prior information to solve lacking large labelled data problem, and the spike model contributes a similar joint and trajectory to coarse refine the test poses by matching, but similarity only provides the approximate location of a joint point, in order to obtain more accurate results, a finely-refine method is proposed, which is introduced in the next section.

3.5 Point distribution based posterior decision multi-network

To achieve a fine refinement of the test results obtained from the JTMPA step in the previous section, conventional data-driven frameworks face challenges due to the scarcity of labeled data. Training or fine-tuning a pose estimation network becomes difficult under such limitations. To address this issue, we propose a point distribution based posterior decision multi-network method. This method decomposes the pose estimation task into a pure classification decision problem, which greatly reduces the dependency on a large volume of labeled data. As a result, we effectively refine the test results without the need for an extensive amount of labeled data.

Algorithm 1 The algorithm of Joint and Trajectory Matching based on the generic one-sided Procrustes Analysis for coarse refinement

```
Input: The spike pose model $\mathbb{M}^S$, and reconstructed 3D poses $\mathbf{X}$ before refinement; // $\mathbb{M}^S$ is obtained from Eq. (8).
Output: Updated 3D poses $\mathbf{X}$ by the proposed coarse refinement method.
for ($f = 0; f < F; f++$) do // Joint Matching: iterate each frame
    $N^J$ = correctJointNumberIndex($\mathbf{X}[f]$); //Only choose the correct joint marked from Section 3.3
    $\mathbf{X}^J[f] \Leftarrow \mathbf{X}[f, N^J, :, :]$; // Test joint matrix
    $\hat{\mathbf{X}}^{SJ} \Leftarrow \mathbb{M}^S[:, :, N^J, :]$; // A set of matrices in the spike model to be matched
    for ($r = 0; r < R; r++$) do
        for ($f_r = 0; f_r < F_r; f_r++$) do
            $\mathbf{X}_{test} = \mathbf{X}^J[f]$;  $\mathbf{X}_{ref} = \hat{\mathbf{X}}^{SJ}[r, f_r]$; // Two matrices to be matched
            $\mathbb{A}, \mathbb{B}$ = genericProcrustesAnalysis($\mathbf{X}_{test}, \mathbf{X}_{ref}$); // Matching
            $E$ = calculateMatchError($\mathbb{A}, \mathbb{B}, \mathbf{X}_{test}, \mathbf{X}_{ref}$); // Evaluate matching error according to Eq. (9)
            if ($E$ is minimum error) then
                $r^J, f_r^J, \mathbb{A}^J, \mathbb{B}^J \Leftarrow r, f_r, \mathbb{A}, \mathbb{B}$; // Best match
            end if
        end for
    end for
    $\mathbf{X}[f]$.update($r^J, f_r^J, \mathbb{A}^J, \mathbb{B}^J, \hat{\mathbf{X}}^{SJ}$); // Update according to Eq. (10)
end for
for ($n = 0; n < N; n++$) do // Trajectory Matching: iterate each joint
    $r^T, \mathbb{A}^T, \mathbb{B}^T$ = findBestMatch($\mathbf{X}^T[n], \hat{\mathbf{X}}^{ST}$); // similar to Joint Matching, find best-matched trajectory according to Eq. (11)
    $\mathbf{X}[n]$.update($r^T, \mathbb{A}^T, \mathbb{B}^T, \hat{\mathbf{X}}^{ST}$); // Update according to Eq. (12)
end for
return $\mathbf{X}$
```

As Fig. 6 shows, A point distribution located around each joint point of the 3D coarse-refined pose gets from Section 3.4 with a to-refine mark to expand the receptive field. For frame f, joint n, the point distribution $\mathbf{X}^D[f, n]$ is represented as follows:

$$\mathbf{X}^D[f, n] \in (\mathbf{X}[f, n] + S \times \mathbf{N}(0, 1)^3)) \in \mathbb{R}^{K^D \times 3} \tag{13}$$

where, S and K^D are hyper-parameters, which means the scale of distribution and number of points, and set as 85 and 200 respectively in our implementation, and $\mathbf{N}(0, 1)$ is the standard normalization distribution function.

From now on, the estimation task is decomposed into a classification decision problem by evaluating each point in distribution from a multi-network γ. In order to improve the stability of the network since the lacking of training data, a multiple network architecture is proposed to eliminate the decision contingency of a single network. The backbone networks are ResNet-34 [17] and VGG-16 [35], and the output of the fully connected layer is a tensor with two elements, which fuse the feature of two backbones. A softmax operation gives a confidence score to the output tensor, which means the possibility of good joint. The input of the multi-network is $w^D \times h^D$ image patches centred at the reprojection point of $\mathbf{X}^D[f, n]$ for the four views, which represents as

$$\mathbf{I}_c^D[f, n] \in \mathbb{R}^{K^D \times w^D \times h^D} \in \mathbf{V}. \tag{14}$$

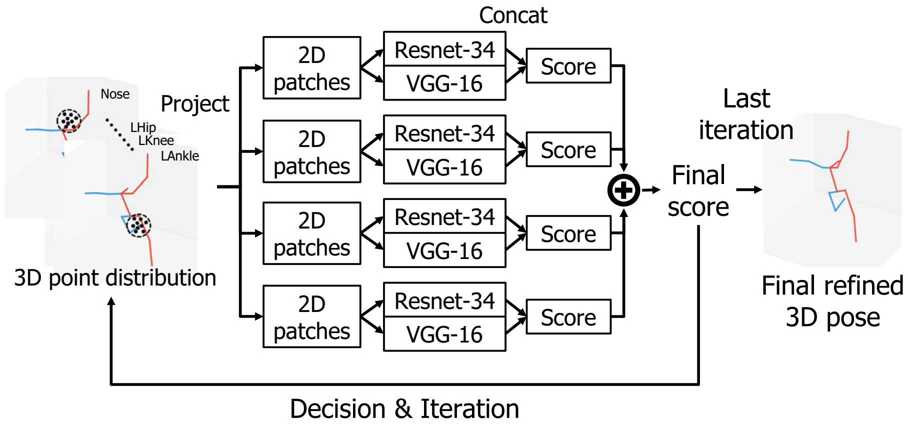


Fig. 6 Concept of point distribution based posterior decision multi-network

where, $w^D = 30$ and $h^D = 30$ in our implementation. Then, the indexes k^D with top-3 confidence score in point distribution is calculated by:

$$\underset{k^D}{\arg\max} \sum_{c=1}^{4} \gamma(\mathbf{I}_c^D[f, n, k^D]). \tag{15}$$

To search for more precise results near high score points, other point distributions are generated with decreased S and K^D in Eq. 13 based on the indexes k^D iteratively. Therefore, in the second iteration, the Eq. 13 is rewritten as following:

$$\mathbf{X}_2^D[f, n] \in (\mathbf{X}^D[f, n, k^D] + S_2 \times \mathbf{N}(0, 1)^3)) \in \mathbb{R}^{K_2^D \times 3} \tag{16}$$

After iterations, the wrong joint points are refined by the highest confidence score, which is the final refined point represented as

$$\mathbf{X}[f, n] \leftarrow \mathbf{X}_{it}^D[f, n, k_{it}^D], \tag{17}$$

where, it is the iteration times, and k_{it}^D is the highest point of it iteration. After do the decision for every joint and frame, the final results is represented as $\mathbf{X^r} = \mathbf{X}$.

4 Experiment result

4.1 Dataset and experimental setting

To evaluate the proposed approach in the wild real competition game scene, as Fig. 7 shows, the dataset video was recorded from the Game of 2014 Japan Inter High School of Men Volleyball with four synchronous camera views located at each corner of the court. In theory, the more camera views, the more accurate the reconstruction results, but due to equipment and shooting license constraints, this paper only discusses the case of four cameras. The video resolution is 4K (3840 × 2160) with 30 frames per second. To focus on spike motion, each sequence is clipped from the start to the end of the spike motion in a whole volleyball game, including four stages of overall spike motion: run up, jump, swing and follow-through. The video size is cropped to remove redundant information.

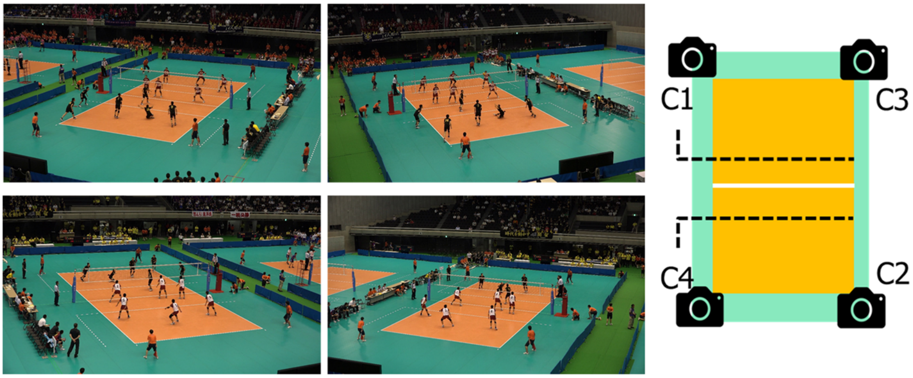


Fig. 7 Four camera views of the dataset videos

To verify the performance of the proposed method, the sequences with different characteristics need to be considered. We analyzed all the spike actions in the whole game and found that the number of blockers affects the estimation accuracy of the spiker's pose. Because the blocker and the spiker are very close together, more blockers mean more occlusion or overlap. To verify the performance of the proposed refinement framework from different aspects, we divided the video sequences of spikes into four categories: no blocker, one blocker, two blockers, and three blockers. And we randomly select 4 sequences in each category as the test dataset. Figure 8 shows examples of each category.

The experimental setting is shown in the following setting: the network part of Section 3.5 runs on a server with one RTX3090 GPU. The training and validation data are 90% and 10% $w^D \times h^D$ patches from the labelled reference sequence of the spike model with a random horizontal flip. The implementation is programmed in Python under the framework of PyTorch. And the other parts of the approach are programmed in Python with the OpenCV library on the PyCharm platform, and run on the i5-7300HQ CPU with 8GB RAM. To achieve data independence, the spike model is constructed by only three typical spike motion sequences from non-test sequences, and the same sequences are used to train the posterior decision multi-network.

4.2 Evaluation metric

Although IoT technology [32–34] is being increasingly used in controllable scenes. By using connected devices such as sensors, and wearable devices, IoT can collect accurate data to obtain the groundtruth of 3D human pose. However, since the dataset videos are taken from the real volleyball game, it is impossible to let players to wearing the sensor. Therefore, the groundtruth of the 3D joint position is reconstructed by the position of four 2D camera views with manual labelling. Following to existing study of the figure skating scene [38], considering the large venue with small target and manual labelling inevitably produce some pixel-level jitter, resulting in millimeter-level error for the 3D joint position in the real world. Therefore, this work utilizes both qualitative and quantitative metrics to evaluate the results.

For the qualitative metric, as Fig. 9 shows, the range of the joint size called the groundtruth range is considered, which is defined that a spherical region in 3D space centered on the joint center. The result is defined as a successful joint if the result joint 3D coordinate is located

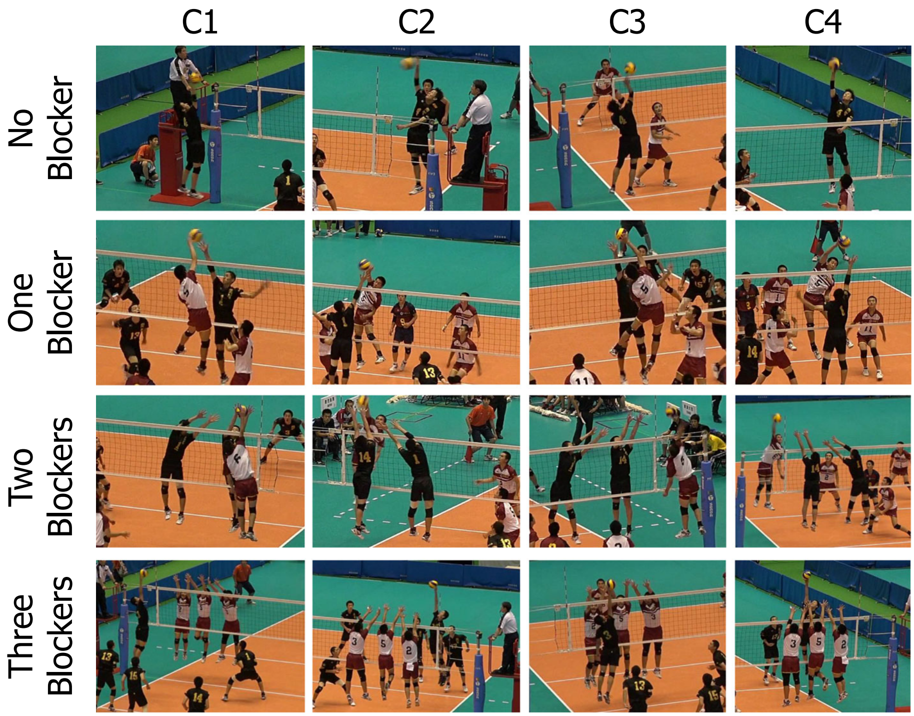


Fig. 8 Examples of spike action categories with different numbers of blockers

in the green circle. This work uses the success rate under different error ranges to evaluate the result, which is defined as follows:

$$SR_{range} = \frac{\#Successful\ Joint}{\#Total\ Joint} \times 100\%. \tag{18}$$

For the quantitative evaluation metric, a common metric the mean per joint position error (MPJPE) is used to evaluate the result, which means the average distance between the ground truth and the result of all joints. Formally, the MPJPE is represented as:

$$\frac{1}{N \times F} \sum_{f=1}^{F} \sum_{n=1}^{N} \left\| \mathbf{X}^{\mathbf{r}}[f, n] - \hat{\mathbf{X}}[f, n] \right\|_1 . \tag{19}$$

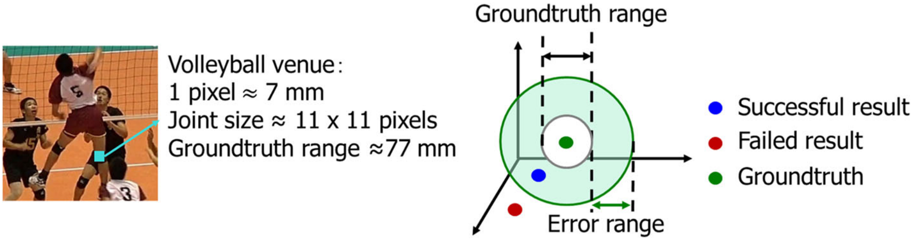


Fig. 9 Groundtruth range and error range

4.3 Comparisons and performance improvement of the state-of-the-art methods

We report the comparison and performance improvement when our proposed refinement framework is applied to the recent state-of-the-art human pose estimation methods. SimCC [23], IntegralPose [37], HRnet [36], DarkNet [44], and OpenPose [4] are used to generate the input pose. To obtain the pose estimation results of the previous methods, we used their released codes and pre-trained models.

Table 2 shows the comparison and performance improvements in success rate under different error ranges when our proposed method is applied to the recent state-of-the-art human pose estimation methods. Our proposed method consistently improves the performance of the state-of-the-art methods. This result indicates that our method not only refines the 2D errors based on the multi-view relationship but also refines the 3D errors based on the prior temporal information. Taking into consideration the fact that the state-of-the-art methods employed in the experiments exhibit structural and learning strategy variations, we argue that our model possesses a high degree of generalizability, thereby enabling its application to other pose estimation methods. Additionally, it is worth noting that our proposed framework necessitates only a limited amount of labeled data, thereby rendering it exceptionally convenient for deployment in real-world competitive scenarios.

4.4 Ablation study

4.4.1 Contribution of different components

To validate the contribution of each component of the proposed approach for refining the pose of the spiker, we perform ablation experiments and analysis on each component. And we use the pretrained OpenPose framework [4], which is a state-of-the-art multi human pose

Table 2 Comparison and performance improvement when our proposed framework is applied to the state-of-the-art methods. SR_{30}, SR_{50}, and SR_{70} mean Success Rate under the 30 mm, 50 mm, and 70 mm error range respectively. MPJPE is the mean per joint position error

Methods	Publication	SR_{30} [%]	SR_{50} [%]	SR_{70} [%]	MPJPE
SimCC [23]	ECCV 2022	50.55	57.61	63.01	119.54 mm
SimCC + Ours	/	64.82 (+14.27)	73.30 (+15.69)	78.53 (+15.52)	82.39 mm (-37.15 mm)
IntegralPose [37]	ECCV 2018	51.01	58.39	64.92	101.50 mm
IntegralPose + Ours	/	64.00 (+12.99)	72.61 (+14.22)	78.97 (+14.05)	78.60 mm (-22.9 mm)
HRnet [36]	CVPR 2019	64.34	68.77	72.70	89.36 mm
HRnet + Ours	/	71.71 (+7.37)	79.67 (+10.90)	84.84 (+12.14)	63.64 mm (-25.72 mm)
DarkNet [44]	CVPR 2020	62.91	67.75	71.67	93.54 mm
DarkNet + Ours	/	70.04 (+7.13)	77.18 (+9.43)	82.86 (+11.19)	64.46 mm (-29.08 mm)
OpenPose [4]	TPAMI 2019	67.12	71.62	75.28	83.19 mm
OpenPose + Ours	/	**76.25 (+9.13)**	**81.89 (+10.27)**	**86.13 (+10.85)**	**56.81 mm (-26.38 mm)**

Note that the bold emphasis means the best performance

Table 3 The success rate of each component in the proposed framework

Method	P1[a]	P2[b]	P3[c]	SR_{30} [%]	SR_{50} [%]	SR_{70} [%]	MPJPE
Baseline	×	×	×	67.12	71.62	75.28	83.19 mm
	√	×	×	74.46 (+7.34)	78.29 (+6.67)	81.59 (+6.31)	70.48 mm (-12.71 mm)
	√	√	×	75.53 (+8.41)	80.67 (+9.05)	84.41 (+9.13)	61.19 mm (-22.00 mm)
Ours	√	√	√	**76.25 (+9.13)**	**81.89 (+10.27)**	**86.13 (+10.85)**	**56.81 mm (-26.38 mm)**

[a] P1: Irrelevant Projection Based Potential Joint Restore
[b] P2: Motion-aware spike model based matching
[c] P3: Point distribution based posterior decision multi-network
Note that the bold emphasis means the best performance

estimation method to generate the input poses. For detail, the experiments are shown as follows:

- Baseline: The baseline is reconstructing multi-view original OpenPose 2D pose to 3D directly without any refinement.
- P1: To validate the contribution of the potential joint restore method, the basic framework combines the multi-view relationship to refine the pose at the 2D level only.
- P1+P2: To validate the contribution of the spike model with matching method, the 3D coarse refinement is added to the experiment #2.
- P1+P2+P3: The overall framework is tested, adding the point distribution based posterior decision multi-network.

Table 3 shows the success rate of the ablation experiments at 30mm, 50mm, and 70mm error ranges. Comparing the success rate of adding P1 or not, the experiment indicates that the result is higher than the result without refinement about 7.34%, 6.67%, and 6.31% at the 30mm, 50mm, and 70mm error ranges respectively after adding the potential joint restore method. The result shows that the proposed method combining the multi-view information is effective to refine the single-view 2D error from the potential joint points. From the result of adding P2, it shows that spike model with matching makes contribution to refine the 3D pose under the condition of lacking labelled data, but the result is based on the similar motion and need to finely refinement. The result of the overall proposed framework shows that the point distribution based posterior decision multi-network provides a more accurate 3D pose. In summary, all the proposed methods in the framework play their due role and achieve good results in the 3D pose refinement of the spiker on the volleyball scene.

Table 4 Effects of joint and trajectory matching strategy in proposed motion-aware spike model based matching method

Method	Joint Matching	Trajectory Matching	SR_{30} [%]	SR_{50} [%]	SR_{70} [%]	MPJPE
	×	×	73.82	79.31	83.58	63.29 mm
	√	×	75.44 (+1.62)	81.33 (+2.02)	85.22 (+1.64)	59.60 mm (-3.69 mm)
	×	√	73.47 (-0.35)	79.00 (-0.31)	82.25 (-1.33)	66.03 mm (+2.74 mm)
Ours	√	√	**76.25 (+2.43)**	**81.89 (+2.58)**	**86.13 (+2.55)**	**56.81 mm (-6.48 mm)**

Note that the bold emphasis means the best performance

4.4.2 Studies on the effect of joint and trajectory matching

To investigate the effects of joint and trajectory matching strategy in the proposed matching-aware spike model based matching method in Section 3.4, we conduct several experiments to evaluate the performance. Table 4 summarizes the performance with different matching methods.

We observe that when only using the joint matching strategy, the performance achieves 2.19 %, 2.55 %, and 1.96 % improvements respectively on the success rate under 30 mm, 50 mm, and 70 mm. Meanwhile, the performance achieves a -3.69 mm decrease on the MPJPE. However, when only using the trajectory matching strategy, the performance has a slight decrease. Specifically, compare with the performance without any matching, the result degrades 0.47 %, 0.39 %, and 1.59 % on the success rate under 30 mm, 50 mm, and 70 mm respectively. On the MPJPE, the performance is worse by a +2.74 mm error. The experimental results illustrate that only using trajectory matching strategy damages the performance. Since the wrong joints crucially affect the trajectory matching at the temporal level. Moreover, these wrong joint points also affect the subsequent multi-network decision for fine-grained refinement.

4.4.3 Studies of the comparison between the single and multiple networks

To investigate the effects of single and multiple networks in the proposed point distribution based posterior decision multi-network method in Section 3.5, we conduct several experiments to evaluate the performance. Table 5 summarizes the performance comparison with single and multiple networks.

We observe that when only using a single VGG16 network, the performance achieves 0.42 %, 0.27 %, and 0.18 % improvements respectively on the success rate under 30 mm, 50 mm, and 70 mm. Meanwhile, the performance achieves a -1.39 mm decrease on the MPJPE. When only using a single ResNet34 network, the performance achieves 0.64 %, 0.87 %, and 1.46 % on the success rate under 30 mm, 50 mm, and 70 mm respectively. On the MPJPE, the performance achieves a -3.84 mm decrease. The experimental results illustrate that the multi-network provides more accurate results than a single network.

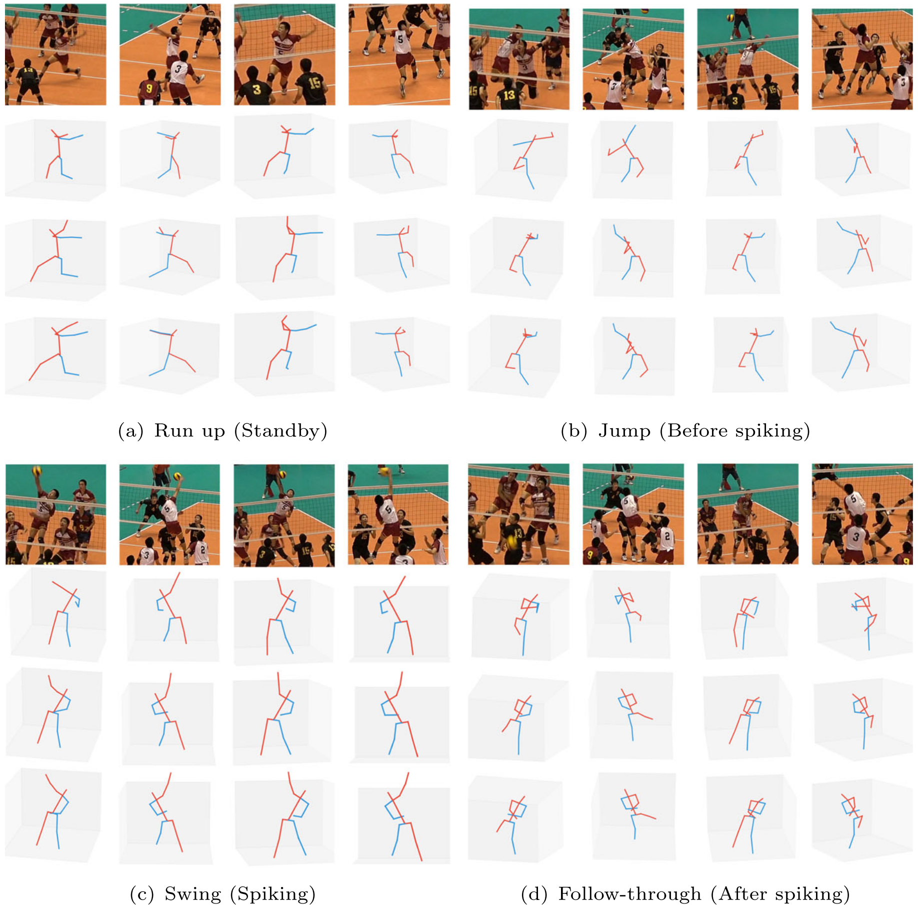
4.5 Visualization

We present a visual representation of the input video frames, baseline 3D pose, refined 3D pose, and ground truth 3D pose in Fig. 10. The comparison clearly illustrates that our proposed method exhibits closer 3D poses with the ground truth when compared to the baseline.

Table 5 Effects of the single and multiple networks in proposed Point distribution based posterior decision multi-network

Method	VGG16	ResNet34	SR_{30} [%]	SR_{50} [%]	SR_{70} [%]	MPJPE
	×	×	75.53	80.67	84.41	61.19 mm
	√	×	75.85 (+0.32)	80.89 (+0.22)	84.56 (+0.15)	59.80 mm (-1.39 mm)
	×	√	76.01 (+0.48)	81.37 (+0.70)	85.64 (+1.23)	57.35 mm (-3.84 mm)
Ours	√	√	**76.25 (+0.72)**	**81.89 (+1.22)**	**86.13 (+1.72)**	**56.81 mm (-4.38 mm)**

Note that the bold emphasis means the best performance



(a) Run up (Standby)

(b) Jump (Before spiking)

(c) Swing (Spiking)

(d) Follow-through (After spiking)

Fig. 10 Visualization of experiment results for each action of spike motion: Each column shows the input cropped frame of four views and its 3D pose observed by four views. First row shows the four view cropped frame of input video; Second row shows reconstructed 3D pose by the baseline conventional pose estimation work [4]; Third row shows the refined 3D pose by proposed approach; Fourth row shows the groundtruth 3D pose reconstructed by manual 2D labelled pose

5 Conclusion

In this paper, the target is 3D pose refinement for refining the estimated pose error work of the spiker on wild real volleyball scene. At the 2D level, to decrease the influence of occlusion and overlap, from the multi-view relationship, this work utilizes the information of other irrelevant camera views, and searches from all potential joint points to refine the 2D pose of a single view. At the 3D level, with the limitation of abnormal pose and lacking labelled data, from the motion similarity, this work proposes a data-independent method, which utilizes the prior motion and temporal information to construct a motion-aware spike model with few labelled sequences for coarse refinement by matching. To finely refinement, a point distribution based posterior decision network is proposed. While expanding the receptive field, the pose estimation task is decomposed into a pure classification decision problem, which greatly reduces the dependence on a large amount of labelled data. The experiment on a wild real multi-view volleyball competition dataset proves the proposed refinement framework, and every component improves the accuracy of the conventional pose estimation work. For future work, since the proposed refinement framework is based on the real volleyball competition, it is expected to be applied in the volleyball analysis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11042-023-16369-8.

Acknowledgements This work has supported by KAKENHI(21K11816).

Data availability The data that support the findings of this study are available from the Japan Volleyball Association (JVA) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the JVA.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Artacho B, Savakis A (2021) Omnipose: a multi-scale framework for multi-person pose estimation. Preprint at http://arxiv.org/abs/2103.10180
2. Askari F, Ramaprasad R, Clark JJ, Levine MD (2022) Interaction classification with key actor detection in multi-person sports videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 3580–3588
3. Bridgeman L, Volino M, Guillemaut J-Y, Hilton A (2019) Multi-person 3D pose estimation and tracking in sports. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp 0–0
4. Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA (2019) Openpose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell
5. Cheng X, Li Z, Du S, Ikenaga T (2020) Body part connection, categorization and occlusion based tracking with correction by temporal positions for volleyball spike height analysis. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences 103(12):1503–1511
6. Cheng K, Zhang Y, He X, Chen W, Cheng J, Lu H (2020) Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 183–192
7. D'Eusanio A, Pini S, Borghi G, Vezzani R, Cucchiara R (2021) Refinet: 3D human pose refinement with depth maps. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp 2320–2327. IEEE
8. Dittakavi B, Bavikadi D, Desai SV, Chakraborty S, Reddy N, Balasubramanian VN, Callepalli B, Sharma A (2022) Pose tutor: an explainable system for pose correction in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 3540–3549
9. Dong J, Jiang W, Huang Q, Bao H, Zhou X (2019) Fast and robust multi-person 3D pose estimation from multiple views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 7792–7801
10. Fang H-S, Xie S, Tai Y-W, Lu C (2017) RMPE: regional multi-person pose estimation. In: ICCV

11. Fieraru M, Khoreva A, Pishchulin L, Schiele B (2018) Learning to refine human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp 205–214
12. Gower JC (1975) Generalized procrustes analysis. Psychometrika 40(1):33–51
13. Guo H, Zou S, Lai C, Zhang H (2021) PHYCOVIS: a visual analytic tool of physical coordination for cheer and dance training. Comput Anim Virtual Worlds 32(1):1975
14. Guo K, Chen T, Ren S, Li N, Hu M, Kang J (2022) Federated learning empowered real-time medical data processing method for smart healthcare. IEEE/ACM Trans Comput Biol Bioinform
15. Guo K, Shen C, Hu B, Hu M, Kui X (2022) RSNet: relation separation network for few-shot similar class recognition. IEEE Trans Multimedia
16. Hartley R, Zisserman A (2003) Multiple view geometry in computer vision. Cambridge University Press
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 770–778
18. Iskakov K, Burkov E, Lempitsky V, Malkov Y (2019) Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp 7718–7727
19. Khan AA, Shaikh AA, Cheikhrouhou O, Laghari AA, Rashid M, Shafiq M, Hamam H (2022) IMG-forensics: multimedia-enabled information hiding investigation using convolutional neural network. IET Image Process 16(11):2854–2862
20. Kocabas M, Karagoz S, Akbas E (2019) Self-supervised learning of 3D human pose using multi-view geometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 1077–1086
21. Li Y, Ji B, Shi X, Zhang J, Kang B, Wang L (2020) Tea: temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 909–918
22. Liu JJ, Newman J, Lee D-J (2020) Body motion analysis for golf swing evaluation. In: International Symposium on Visual Computing. Springer, pp 566–577
23. Li Y, Yang S, Liu P, Zhang S, Wang Y, Wang Z, Yang W, Xia S-T (2022) SIMCC: a simple coordinate classification perspective for human pose estimation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. Springer, pp 89–106
24. Mei J, Chen X, Wang C, Yuille A, Lan X, Zeng W (2019) Learning to refine 3D human pose sequences. In: 2019 International Conference on 3D Vision (3DV). IEEE, pp 358–366
25. Moon G, Chang JY, Lee KM (2019) Posefix: model-agnostic general human pose refinement network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 7773–7781
26. Muhammad K, Ahmad J, Rho S, Baik SW (2017) Image steganography for authenticity of visual contents in social networks. Multimed Tools Appl 76:18985–19004
27. Mukherjee S, Sanyal G (2020) Image steganography with n-puzzle encryption. Multimed Tools Appl 79(39–40):29951–29975
28. Napolitano S, Perciavalle V, Ascione A (2017) Pilot study in youth volleyball: Video analysis as a didactic tool. Giornale Italiano di Educazione alla Salute, Sport e Didattica Inclusiva 1(2)
29. Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Harvesting multiple views for marker-less 3D human pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 6988–6997
30. Qiu H, Wang C, Wang J, Wang N, Zeng W (2019) Cross view fusion for 3D human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp 4342–4351
31. Shafiq M, Gu Z (2022) Deep residual learning for image recognition: a survey. Appl Sci 12(18):8972
32. Shafiq M, Tian Z, Bashir AK, Du X, Guizani M (2020) IoT malicious traffic identification using wrapper-based feature selection mechanisms. Comput Secur 94:101863
33. Shafiq M, Tian Z, Sun Y, Du X, Guizani M (2020) Selection of effective machine learning algorithm and bot-IoT attacks traffic identification for internet of things in smart city. Futur Gener Comput Syst 107:433–442
34. Shafiq M, Gu Z, Cheikhrouhou O, Alhakami W, Hamam H (2022) The rise “internet of things”: review and open research issues related to detection and prevention of IoT-based security attacks. Wirel Commun Mob Comput 2022:1–12
35. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Preprint at http://arxiv.org/abs/1409.1556
36. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 5693–5703
37. Sun X, Xiao B, Wei F, Liang S, Wei Y (2018) Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV). pp 529–545

38. Tian L, Cheng X, Honda M, Ikenaga T (2022) Multi-view 3D human pose reconstruction based on spatial confidence point group for jump analysis in figure skating. Complex Intell Syst 1–15
39. Véges M, Lőrincz A (2020) Temporal smoothing for 3d human pose estimation and localization for occluded people. In: International Conference on Neural Information Processing. Springer, pp 557–568
40. Wang J, Qiu K, Peng H, Fu J, Zhu J (2019) AI coach: deep human pose estimation and analysis for personalized athletic training assistance. In: Proceedings of the 27th ACM International Conference on Multimedia. pp 374–382
41. Wang C, Qiu H, Yuille AL, Zeng W (2019) Learning basis representation to refine 3D human pose estimations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33. pp 8925–8932
42. Yang C, Xu Y, Shi J, Dai B, Zhou B (2020) Temporal pyramid network for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 591–600
43. Zeng A, Yang L, Ju X, Li J, Wang J, Xu Q (2022) Smoothnet: a plug-and-play network for refining human poses in videos. In: European Conference on Computer Vision. Springer
44. Zhang F, Zhu X, Dai H, Ye M, Zhu C (2020) Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 7093–7102
45. Zhao H, Jia J, Koltun V (2020) Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 10076–10085
46. Zhou C, Ren Z, Hua G (2020) Temporal keypoint matching and refinement network for pose estimation and tracking. In: European Conference on Computer Vision. Springer, pp 680–695
47. Zhu K, Wong A, McPhee J (2022) Fencenet: fine-grained footwork recognition in fencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 3589–3598
48. Zou J, Li B, Wang L, Li Y, Li X, Lei R, Sun S (2018) Intelligent fitness trainer system based on human pose estimation. In: International Conference On Signal and Information Processing, Networking and Computers. Springer, pp 593–599

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yanchao Liu[1] · **Xina Cheng**[2] · **Takeshi Ikenaga**[1]

Xina Cheng
xncheng@xidian.edu.cn

Takeshi Ikenaga
ikenaga@waseda.jp

[1] Graduate School of Information, Product and Systems, Waseda University, 2-7, Kitakyushu 8080135, Fukuoka, Japan

[2] School of Artificial Intelligence, Xidian University, No. 2 South Taibai Road, Xi'an 710126, Shaanxi, China