Synergetic Reconstruction from 2D Pose and 3D Motion for Wide-Space Multi-Person Video Motion Capture in the Wild

Takuya Ohashi[1,2] Yosuke Ikegami[2] Yoshihiko Nakamura[2]
[1]NTT DOCOMO [2]The University of Tokyo
takuya.ohashi.ht@nttdocomo.com {ikegami,nakamura}@ynl.t.u-tokyo.ac.jp

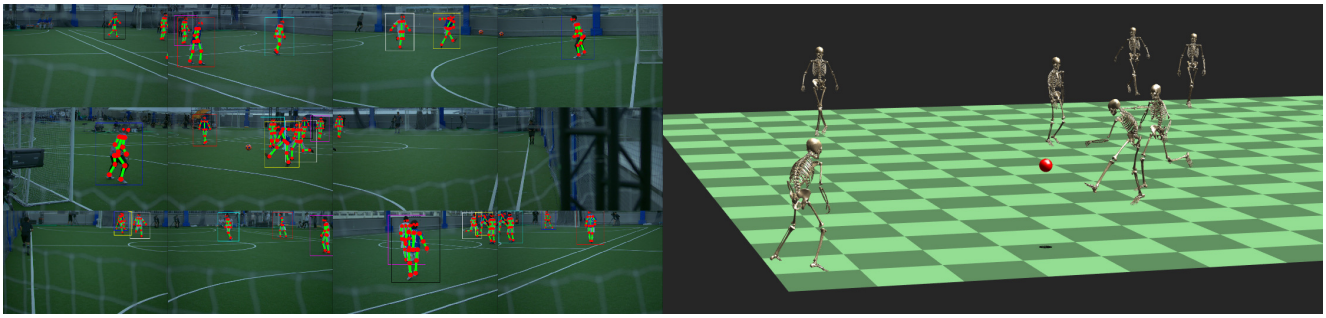


Figure 1: All futsal players' motions were captured using 12 video cameras surrounding the court. (left) Input images and reprojected joint position. (right) Bone CG drawing based on the calculated joint angles.

Abstract

Although many studies have been made on markerless motion capture, it has not been applied to real sports or concerts. In this paper, we propose a markerless motion capture method with spatiotemporal accuracy and smoothness from multiple cameras, even in wide and multi-person environments. The key idea is predicting each person's 3D pose and determining the bounding box of multi-camera images small enough. This prediction and spatiotemporal filtering based on human skeletal structure eases 3D reconstruction of the person and yields accuracy. The accurate 3D reconstruction is then used to predict the bounding box of each camera image in the next frame. This is a feedback from 3D motion to 2D pose, and provides a synergetic effect to the total performance of video motion capture. We demonstrate the method using various datasets and a real sports field. The experimental results show the mean per joint position error was 31.6mm and the percentage of correct parts was 99.3% under five people moving dynamically, with satisfying the range of motion. Video demonstration, datasets, and additional materials are posted on our project page[1].

1. Introduction

Human motion data are used widely in various fields such as sports training, CG production, rehabilitation, medical diagnosis, behavioral understanding, and even humanoid robot operation [41, 30, 35]. To obtain the data, various motion capture methods are developed such as optical motion capture, by which reflective markers are attached to characteristic parts of the body; then these 3D positions are measured [1, 5]. Inertial motion capture uses IMU sensors attached to body parts; then positions are calculated using sensor speed [6, 2]. Markerless motion capture uses a depth camera or single/multiple RGB video cameras [32, 36, 3, 4]. Nevertheless, although various means of using motion data exist, motion capture is used only in limited locations. Few examples have been reported of motion capture being conducted in places that have apparently high value, such as sports matches, concerts, and on streets.

Why are motion data not captured in the real world? In actuality, motion capture in real world conditions is challenging. Because human motion is continuous, the motion data must also be continuous. The real world has three specific factors that make capture difficult. The first is multiple persons. Multiple subjects cause occlusion, and require identification and tracking. The second difficulty is from the large measurement field. The wider a measurement field, the greater the calibration error. Precise calibration is required. Furthermore, the measurement field is sometimes open, which means some persons going out of the field and the others coming into the field. The third difficulty derives from the real environment, namely, non-ideal and restricted measurement conditions. For competitive events or concerts, it is required to avoid any constraints for the

[1] http://www.ynl.t.u-tokyo.ac.jp/research/vmocap-syn

measurement, such as markers, IMU sensors, or specific shirts/pants. Furthermore, the other constraints exist such as being forced to take measurements in a severe lighting condition or being unable to set the sensor at desired position. Because of these various difficulties, even with the latest technology, motion capture in the real world has not been fully developed.

In this paper, we discuss the multi-person video motion capture, which means image-based 3D human motion reconstruction with spatiotemporal accuracy and smoothness even in a challenging multi-person environment, by extending the single-person video motion capture method [31]. We use synchronized multiple calibrated cameras to record video images of human subjects from different directions. We also use a human skeletal model for reconstructing 3D motion by spatiotemporal filtering of joint movements. Our key idea is to predict each person's 3D pose and determine the bounding box small enough. Using the bounding box, the keypoint positions of each subject in each image are estimated using top-down pose estimation approach [40, 34]. They are received as part confidence maps (PCM). Using the PCM of multi-camera images and a predicted 3D pose, probable keypoint positions can be calculated. By minimizing the position errors and the skeletal model's corresponding joint positions, the skeletal model's 3D pose is reconstructed. The 3D reconstructed motion is then used to predict the bounding box of each camera image in the next frame. This feedback from 3D motion reconstruction to 2D pose estimation provides the synergetic effect to the total performance of video motion capture.

The proposed method is tested with various datasets [10][1]. Thereby, the performance is evaluated quantitatively. We also apply the proposed method to actual futsal matches and verify it in real environments. Additionally, because the proposed method uses inverse kinematics (IK) for optimization, it is possible to calculate not only the position but also the joint angle considering the range of motion (RoM). As a qualitative evaluation, bone CG is drawn using the joint angle as Figure 1.

2. Related work

2.1. Single-view pose estimation

Human 2D pose estimation from a single image is a task of detecting human keypoint positions in an image such as knees and shoulders. Typically, two approaches are used. A top-down approach first detects the positions of multiple people in an image as a bounding box; it then estimates the keypoint positions of the single person in the cropped image [21, 15, 40, 34]. A bottom-up approach first estimates 2D keypoint positions of all persons from the entire image. It then associates them for each person [38, 14, 24]. In general, a top-down approach is more accurate. A bottom-up approach is faster. However, a top-down approach relies heavily on human detection results for accuracy. Therefore, the estimation is likely to fail in severe occlusion environments.

In recent years, some studies have estimated human 3D poses solely from a single image, by extending detected 2D keypoint positions to 3D spaces [29, 26, 8], directly estimating 3D poses [27, 12, 23, 45], or estimating not only poses but also detailed body shapes [39, 19]. However, 3D pose estimation from a single image is fundamentally an ill-posed problem. Various assumptions must be set. Therefore, the estimation accuracy obtained under a complex environment, such as a multi-person environment, is markedly inferior to methods using multiple cameras.

2.2. Multi-view 3D pose estimation

Work examining 3D pose estimation using multiple cameras has been reported widely. Most early research efforts extracted a person region from an image, considered the region of the human body in 3D space, and tracked the region continuously over time [16, 33]. This tracking-based approach can estimate motion independently of the subject's pose. It has achieved remarkable results. However, for preparation, it is necessary to create a detailed human model including clothing. The approach might therefore fail according to light conditions, backgrounds, and the clothing of the subject.

In recent years, as 2D pose estimation methods have achieved remarkable results, approaches combining 2D pose estimation and multi-view geometry have been assessed actively, such as reconstructing estimated keypoints in 3D [22, 17, 13] or comparing the 3D keypoints probability and 3D pictorial structure [10, 11, 18]. However, most of these work take an approach of not reconstructing limb when its 2D pose estimation becomes difficult. As a result, continuity, which is essential for the motion capture, is lost.

One earlier study [31] proposed a method that uses a bottom-up approach [38, 14] from multiple cameras to estimate 3D keypoint positions, and applies filtering based on the continuity of skeletal structure and motion to the positions, and realized high-accuracy and smooth motion capture only from a few cameras. However, this approach presented three difficulties. First, because the approach specifically examines a single person, if there are multiple persons, the likelihood of keypoint positions cannot be computed. Second, the measurement area is narrow because the area is mainly limited to an overlapping area of four cameras' respective fields of view. Third, although IK computation is used for the filtering, the RoM is not considered. As a result, strange poses might be reconstructed. In this paper, we resolve these difficulties and propose a method for high-accuracy and smooth motion capture with satisfying RoM, even under multi-person and wide area environment.

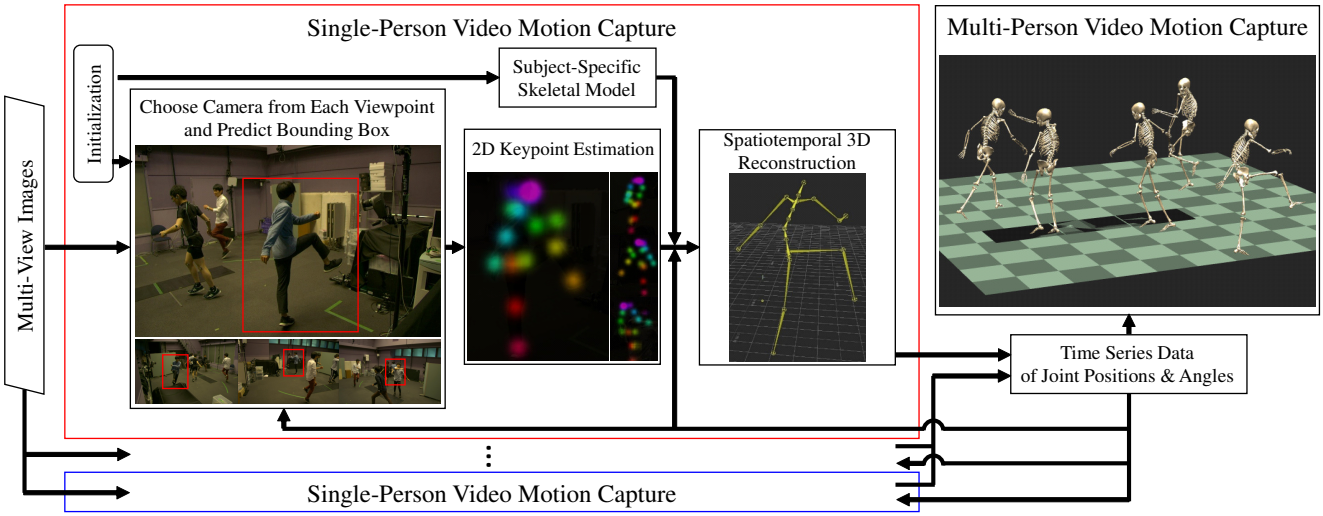


Figure 2: Flowchart of the multi-person video motion capture

3. Synergetic reconstruction

The proposed 3D motion reconstruction is performed using n_c synchronized calibrated cameras placed around n_p subjects. In measurements, to avoid difficulties by which the subject becomes invisible to one camera, multiple cameras having different fields of view are set at one location. We designate this location a viewpoint and assume n_v as the number of viewpoints, $\mathbb{C}_v$ as the set of cameras placed at viewpoint v, and $n_{\mathbb{C}_v}$ as the number of cameras at v.

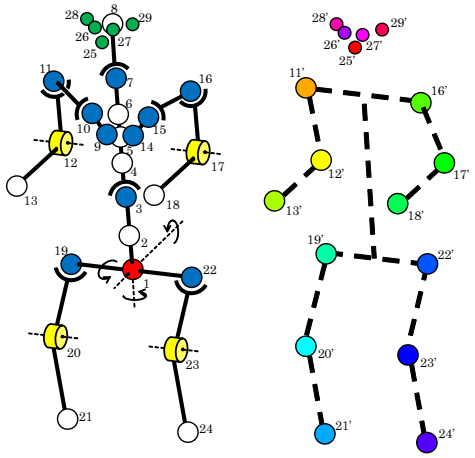
$$n_c = \sum_{v}^{n_v} n_{\mathbb{C}_v} \tag{1}$$

A flowchart showing the proposed method is presented in Figure 2. Each subject's keypoint positions are estimated using a top-down pose estimator, HRNet [40, 34]. The data are received as PCM. We use PCM instead of the pixel location of the keypoint. Using PCM, we perform spatiotemporal optimization of human skeletal model and reconstruct the 3D motion. The skeletal model represents a virtual open tree-structure kinematic chain with 40 degrees of freedom (DoF), as depicted in Figure 3[a]. Then, the 3D pose in the next time frame is calculated accurately. The pose is passed to the HRNet as bounding box information. By applying the process above in parallel for each subject and by repeating the process for each time frame continually, multi-person video motion capture is realized.

Although camera calibration and system initialization are important for implementing the proposed method, they are not the main topics. Therefore, we describe details in the *Appendix* and use μ_i, which represents perspective projection transformation to camera i.

3.1. Determine bounding box from 3D motion

In recent years, top-down pose estimation approaches have achieved remarkable results. If a suitable bounding box is specified, the estimator can compute only the intended person's PCM robustly and accurately, even in severe occlusion environments, as shown in Figure 4. However, pose estimation in a multi-person environment is still challenging. One factor is that a suitable person region cannot be detected (e.g., a wrist or an ankle is cut).



[a] Skeletal Model [b] Keypoint Positions

Figure 3: Correspondence of a human skeletal model and keypoint positions. In [a], joint color indicates as 6DoF (red), 3DoF (blue), and 1DoF (yellow).

The proposed method accomplishes high-accuracy motion capture. If the frame rate is moderately high, then the subject's current 3D pose can be predicted from the calculated past 3D motion. Also, the bounding box position is calculable using perspective projection transformation. The calculation cost is slight. Therefore, we adopt a state-of-the-art top-down pose estimation approach: HRNet. The human

Figure 4: 2D keypoint estimation using HRNet [40, 34]. By specifying a bounding box for the target person, the intended person's specific PCM can be estimated. The input image is from the OCHuman Dataset [43].

region is determined from past 3D motion. The bounding box is calculated simply as shown below.

$$ {}^{t+1}_{l}\mathbf{B}_i = \begin{bmatrix} \{\max(\left[\mu_i({}^{t+1}_{l}\mathbf{P}_{pred})\right]_x) + \min(\left[\mu_i({}^{t+1}_{l}\mathbf{P}_{pred})\right]_x)\}/2 \\ \{\max(\left[\mu_i({}^{t+1}_{l}\mathbf{P}_{pred})\right]_y) + \min(\left[\mu_i({}^{t+1}_{l}\mathbf{P}_{pred})\right]_y)\}/2 \\ m\{\max(\left[\mu_i({}^{t+1}_{l}\mathbf{P}_{pred})\right]_x) - \min(\left[\mu_i({}^{t+1}_{l}\mathbf{P}_{pred})\right]_x)\} \\ m\{\max(\left[\mu_i({}^{t+1}_{l}\mathbf{P}_{pred})\right]_y) - \min(\left[\mu_i({}^{t+1}_{l}\mathbf{P}_{pred})\right]_y)\} \end{bmatrix} \quad (2) $$

$$ {}^{t+1}_{l}\mathbf{P}_{pred} = \frac{3}{2}\,{}^{t}_{l}\mathbf{P} - {}^{t-1}_{l}\mathbf{P} + \frac{1}{2}\,{}^{t-2}_{l}\mathbf{P} \quad (3) $$

Therein, ${}^{t+1}_{l}\mathbf{B}_i$ represents the predicted center position and size of the bounding box of person l at time $t+1$ on camera i, ${}^{t}_{l}\mathbf{P}$ represents 3D positions of all joints, and m represents a constant positive value whole body becomes just visible. All joints mean $n_j = 29$ joints, as depicted in Figure 3[a]. Note that, assuming uniformly accelerated motion, the future 3D pose is calculated as ${}^{t+1}\mathbf{P} = 2\,{}^{t}\mathbf{P} - 2\,{}^{t-1}\mathbf{P} + {}^{t-2}\mathbf{P}$. However, we use ${}^{t+1}\mathbf{P}_{pred} = ({}^{t+1}\mathbf{P} + {}^{t}\mathbf{P})/2$ as the predicted 3D pose.

For the proposed method, we use a pretrained model of HRNet that has been trained on the COCO dataset [25]. The input image is resized and trimmed to the size of $W' \times H' \times 3$ according to the bounding box. PCM is computed from a cropped image. The size of the cropped image is fixed ($W' = 288$, $H' = 384$). The number of keypoints is $n_k = 17$, which consists of 12 joints (shoulders, elbows, wrists, hips, knees, and ankles) and 5 feature points (eyes, ears, and nose) as Figure 3[b].

In addition, because HRNet was trained by assuming that the body is not tilted much, the estimation might fail when the body is tilted greatly with respect to the image vertical direction, such as during a handstand or cartwheel. With the proposed method, by rotating the bounding box, one can estimate the PCM correctly. The rotation angle is derived from the inclination of predicted vector connecting the torso and neck.

$$ {}^{t+1}_{l}\mathbf{B}'_i = \frac{\pi}{2} - \mathrm{atan2}\Big(\left[\mu_i({}^{t+1}_{l}P^{n(1)}_{pred})\right]_y - \left[\mu_i({}^{t+1}_{l}P^{n(6)}_{pred})\right]_y , \left[\mu_i({}^{t+1}_{l}P^{n(1)}_{pred})\right]_x - \left[\mu_i({}^{t+1}_{l}P^{n(6)}_{pred})\right]_x \Big) \quad (4) $$

In this equation, n represents the joint position of the human skeletal model. The number represents the specific position as depicted in Figure 3[a]. Only 11 keypoints (shoulders, elbows, wrists, eyes, ears, and nose) are calculated from the rotated bounding box.

Also, because multiple cameras with different fields of view are set at one viewpoint, the camera to which the target person is most visible should be chosen for 2D keypoint estimation at each viewpoint. For the proposed method, this choice is performed by the predicted joint position as explained before.

$$ i(v,t,l) = \arg\min_{i \in \mathbb{C}_v} \Big\{ \Big(\left[\mu_i({}^{t+1}_{l}P^{n(1)}_{pred})\right]_x - \frac{I_x}{2} \Big)^2 + \Big(\left[\mu_i({}^{t+1}_{l}P^{n(1)}_{pred})\right]_y - \frac{I_y}{2} \Big)^2 \Big\} \quad (5) $$

Therein, I stands for the camera image resolution.

3.2. Spatiotemporal 3D motion reconstruction

For obtaining the 3D keypoint position, 3D reconstruction of the detected 2D keypoint position by multiple cameras is conceivable, but this simple method might fail in a severe occlusion environment because of false and missing detection. However, even when the keypoint position is detected erroneously, the PCM might indicate the probability at the correct keypoint position. For example, in Figure 4, the PCM of the left ankle of the left person shows the probability at both incorrect and correct positions. In other words, PCM is a stochastic field that includes both true positive results (TP) and false positive results (FP). If only TP is referred successfully, then robust 3D reconstruction can be achieved even in a severe occlusion environment.

One can consider a lattice space ${}^{t+1}_{l}\mathbb{L}^n$ with ${}^{t+1}_{l}P^n_{pred}$ as a center, s as the interval, and ${}^{t+1}_{l}L^n_{a,b,c}$ as one point of it.

$$ {}^{t+1}_{l}\mathbb{L}^n := \left\{ {}^{t+1}_{l}P^n_{pred} + s \begin{bmatrix} a \\ b \\ c \end{bmatrix} \;\middle|\; -k \le a,b,c \le k \right\} \quad (6) $$

k : constant positive integers

a, b, c : integers

$$ {}^{t+1}_{l}L^n_{a,b,c} \in {}^{t+1}_{l}\mathbb{L}^n \quad (7) $$

Using perspective projection transformation, one can obtain the PCM value of arbitrary 3D point at camera i. Considered simply, if ${}^{t+1}_{l}P^n_{pred}$ is predicted accurately, the most probable keypoint position is one point of this grid where the sum of PCM value is maximum. This calculation is robust against large false estimation. It is lighter than considering a huge stochastic field by projecting multiple PCMs into 3D.

However, the proposed method aims for a multi-person environment. The top-down approach attempts to compute the PCM of the intended person in the bounding box, but it has limitations. It might compute unintended PCM if truly severe occlusion occurs, as presented in Figure 5. However, the PCM computation behavior in such an occlusion

environment is difficult to treat quantitatively. Even under similar environments, various estimation results can be obtained, such as false, mixed, and ideal estimation.

One option is not to refer to PCM in such an occlusion environment, but this approach disregards TP that might be present in the PCM. In the proposed method, we assume that the reliability of PCM decreases in occlusion environment. We assign a constant weight to the PCM. The most probable keypoint position is acquired as

$$ {}^{t+1}_{l}P^{n}_{key} = \underset{-k\leq a,b,c\leq k}{\arg\max} \sum_{v}^{n_v} {}^{t+1}_{l}w^{n}_{i}\, {}^{t+1}_{l}\mathcal{S}^{n}_{i}(\mu_i({}^{t+1}_{l}L^{n}_{a,b,c})) \quad (8) $$

$$ {}^{t+1}_{l}w^{n}_{i} = \begin{cases} g & \text{if } \mu_i({}^{t+1}_{l}P^{n}_{pred}) \text{ is occluded by other } \mu_i({}^{t+1}P^{n}_{pred}) \\ 1 & \text{otherwise.} \end{cases}, \quad (9) $$

where ${}^{t+1}_{l}\mathcal{S}^{n}_{i}(X)$ represents a function for obtaining the PCM value on camera i at time $t+1$ of joint n of person l. Also, g represents a constant value in (0,1).

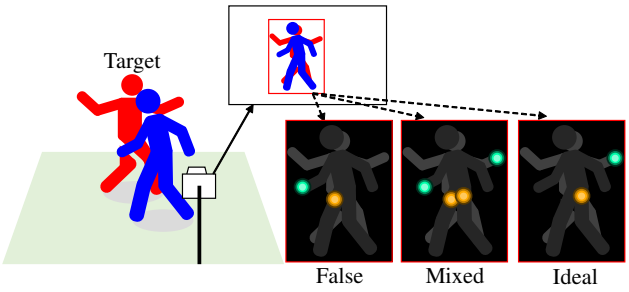


Figure 5: PCM computation with truly severe occlusion. The pose estimator estimates the keypoints of the right wrist and right hip of the red person, but the actual results are not known.

Next, by referencing the calculated keypoint position, we compute the joint position of the skeletal model. By the proposed method, the skeletal model's joint angle is optimized using IK [9] by the keypoint position as the target position, while referencing the correspondence depicted in Figure 3.

$$ {}^{t+1}_{l}\mathbf{Q} = \arg\min \sum_{n}^{n_k} \frac{1}{2}\, {}^{t+1}_{l}W^{n} || {}^{t+1}_{l}P^{n}_{key} - {}^{t+1}_{l}P^{n} ||^2 \quad (10) $$

$$ s.t. \quad {}^{t+1}_{l}\dot{\mathbf{P}} = {}_{l}\mathbf{J}\ {}^{t+1}_{l}\dot{\mathbf{Q}} \quad (11) $$

$$ {}^{t+1}_{l}W^{n} = \sum_{v}^{n_v} {}^{t+1}_{l}\mathcal{S}^{n}_{i}(\mu_i({}^{t+1}_{l}P^{n}_{key})) \quad (12) $$

Therein, ${}^{t+1}_{l}\mathbf{Q}$ represents the joint angle of the person l at time $t+1$. Also, ${}_{l}\mathbf{J}$ represents the Jacobian matrix.

Although the joint positions can be computed by the optimization above, these positions do not incorporate consideration of the temporal continuity of the motion. To obtain the smooth motion, the joint position is smoothed using a low-pass filter $\mathcal{F}$ consisting of time series data of joint positions.

$$ {}^{t+1}_{l}\mathbf{P}_{smo} = {}^{t+1}_{l}\mathcal{F}({}^{t+1}_{l}\mathbf{P}) \quad (13) $$

However, when the smoothing above is performed, the skeletal structure is collapsed. Spatial continuity is lost. In addition, although only the link length is considered in the above IK computation, each joint angle is expected not to deviate from RoM. Then, the skeletal model is optimized using IK again by the smoothed joint position as the target position as

$$ {}^{t+1}_{l}\mathbf{Q}' = \arg\min \sum_{n}^{n_k} \frac{1}{2}\, || {}^{t+1}_{l}P^{n}_{smo} - {}^{t+1}_{l}P'^{n} ||^2 \quad (14) $$

$$ s.t. \quad {}^{t+1}_{l}\dot{\mathbf{P}}' = {}_{l}\mathbf{J}\ {}^{t+1}_{l}\dot{\mathbf{Q}}' \quad (15) $$

$$ \mathbf{Q}^{-} \leq {}^{t+1}_{l}\mathbf{Q}' \leq \mathbf{Q}^{+} \quad (16) $$

where $\mathbf{Q}^{-}$ and $\mathbf{Q}^{+}$ represent the minimum and maximum values of RoM [42]. By the computation above, joint positions and angles with spatiotemporal accuracy are acquired.

By repeatedly computing the above processes, single-person motion capture is achieved. By processing in parallel to the number of subjects, multi-person video motion capture can be realized.

4. Experimental results

The proposed method is applied to various datasets as shown in Table 1, including an original one called Studio. For evaluation, various metrics are proposed. In this paper, we use the percentage of correct parts (PCP), percentage of correct keypoints (PCK), and mean per joint position error (MPJPE). With PCP, a limb is considered detected if the distance between the two calculated joint positions and true limb joint positions is less than half of the limb length. With PCK, a calculated joint is considered correct if the distance between the calculated and the true joint is within a certain threshold. MPJPE denotes the average distance between the calculated and true joint positions.

Dataset	n_c	n_v	n_p	I	F	M
Shelf [10]	5	5	2-4	1032×776	20	3×3
Studio[1]	8	4	1-5	1920×1200	60	5×7
Futsal	12	4	7-8	1920×1200	60	16×24

Table 1: Dataset overview. n_c: number of cameras, n_v: number of viewpoints, n_p: number of persons, I: image resolution, F: frame rate, M: approximate measurement filed size [m].

In the bone CG depicted in the following figure, the bone length is different for each subject. Also, the motion is updated according to the calculated joint angle.

4.1. Evaluation using public dataset

The proposed method is applied to a public dataset called Shelf [10], in which four people are mutually interacting.

They are recorded by five cameras (Figure 6). We follow the same evaluation metrics as those used in earlier work [11, 18, 17, 13]; we also use PCP for evaluation.

A few points are noteworthy. First, because some subjects are invisible in the initial frame, it is impossible to calculate their initial joint position and link lengths. Therefore, these subjects are excluded from analyses. Only the subjects who can be initialized are reconstructed. Second, the provided ground truth keypoints and our skeletal models' joints differ. Therefore, only body parts, except for the head, are used to calculate PCP. Third, to calculate PCP, alternative ways are proposed [17, 13], using the average of the distance of the two joints. Therefore, we calculate PCP using two methods. The results are presented in Table 2.

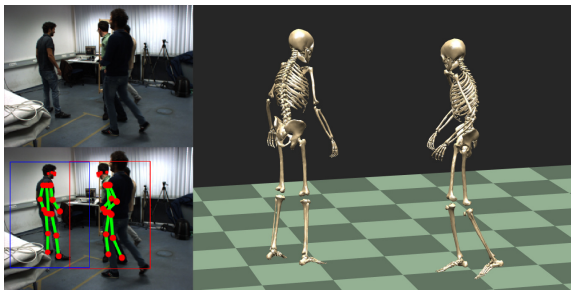


Figure 6: Qualitative result on the Shelf dataset [10]

Method	Actor 1	Actor 2	Actor 3
Belagiannis *et al.* [11]	75.3	69.7	87.6
Ershadi-Nasab *et al.* [18]	93.3	75.9	94.8
Bridgeman *et al.* [13]	**98.8**	85.9	**97.1**
Ours	98.4	-	**97.1**
Dong *et al.* [17]	98.8	94.1	97.8
Bridgeman *et al.* [13]	99.7	92.8	97.7
Ours	**99.9**	-	**97.9**

Table 2: Comparison of PCP to the Shelf dataset [10]. Upper part is calculated from the two joint positions constituting the limb, and lower is calculated from the average.

Results demonstrate that the proposed method can reconstruct 3D motion robustly and accurately, even in a multi-person environment. Moreover, the method can achieve equal or better performance to that obtained from existing methods.

However, in the dataset, the subject motions are slow and slight. It is questionable whether this accuracy can be trusted when used in actual sports scenes. Therefore, to examine specific problems such as dynamic motion, complex poses, and multiple persons, we create an original dataset measuring the multiple subjects, and do evaluation.

4.2. Evaluation using original dataset

Using eight RGB cameras (acA1920-155uc; Basler AG) at 60 Hz, 15 subjects were recorded. Also, using 17 infrared cameras (Eagle and Raptor-4l; Motion Analysis Corp.) at 200 Hz, two subjects with 44 reflective markers were measured synchronously. For this measurement, two cameras were set at each viewpoint to cover the entire measurement field. Eight motions were measured such as dances, boxing, and handstands, as shown in Figure 7. Datasets will be published together with the camera parameters and the marker positions for related research[1]. The results are presented in Table 3.

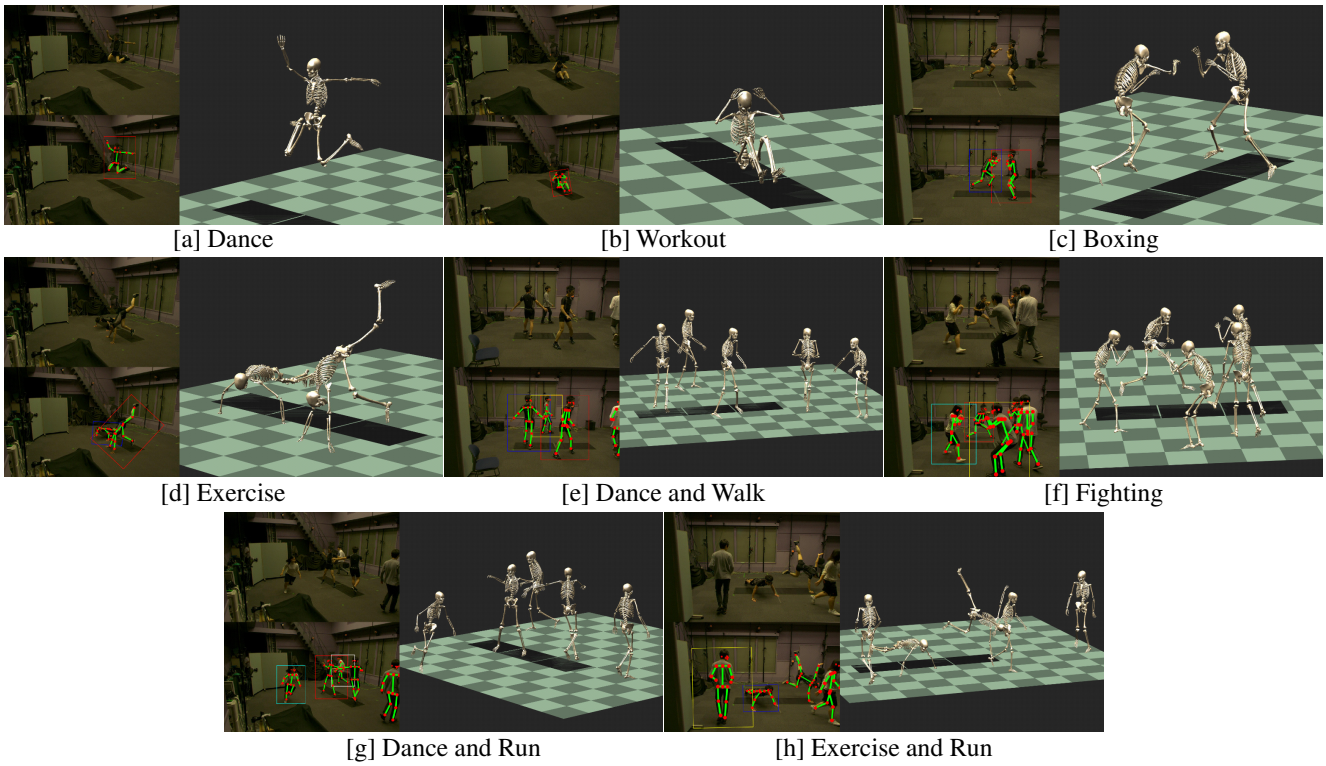
Results demonstrate that the proposed method can reconstruct 3D motion robustly and accurately, even in a five-person environment as well as in a single-person environment. The proposed method achieved 31.7mm in MPJPE and 99.3% in PCP under five-person dynamic movement. The results indicate that the proposed method achieved almost equal or better performance under a single-person environment than earlier reported methods [31] (26.1mm in MPJPE and 95.8% in PCK@50 mm without RoM). Furthermore, even in a challenging environment in which the human pose is greatly inclined, such as a handstand or push-up, which are generally difficult in pose estimation, the 3D motion can be acquired by rotating the bounding box, and achieve over 95.0% in PCP under five-person envirinment.

The bone CG shows that the proposed RoM restriction works even under dynamic motion, thereby preventing strange pose reconstruction. However, this restriction sometimes has an adverse effect: when performing a dynamic motion such as swinging the arms, the optimization might fall into a singular posture and become unable to acquire optimal joint positions. Comparing the results obtained with and without RoM reveals that the latter achieves higher accuracy. Therefore, if only 3D joint positions are needed, the RoM is not expected to be restricted. However, if the motion data are used for CG production or medical diagnosis etc., then 3D reconstruction with RoM is more suitable.

4.3. Experiment on futsal field

To verify the proposed method in an actual environment, we measured futsal games. In the measurement, to cover about two-thirds of the court with the camera's field of view, 12 RGB cameras were set at four corners; eight players were recorded. The futsal ball was detected by color from each camera. It was reconstructed in 3D. As an aside, using the ball trajectories, bundle adjustment [37] was performed, and camera parameters were acquired. The results are presented in Figure 8.

Because no ground truth is available, the results are only qualitative evaluation, but the results of reprojected joint positions onto the input image and the bone CG show that motion capture can be achieved with high accuracy almost equal to that of experiments on earlier datasets. Using only a few cameras, all players' detailed motion was successfully acquired.



[a] Dance [b] Workout [c] Boxing

[d] Exercise [e] Dance and Walk [f] Fighting

[g] Dance and Run [h] Exercise and Run

Figure 7: Qualitative results on the Studio dataset[1]

Dataset	[a]	[b]	[c]		[d]		[e]		[f]		[g]		[h]		[e,f,g]
Number of Persons	1	1	2		2		5		5		5		5		5
Total Time[s]	35.2	32.8	28.2		30.8		33.8		30.9		34.2		30.1		98.9
Need Rotation?	No	Yes	No		Yes		No		No		No		Yes		No
Actor ID	1	1	1	2	1	2	1	2	1	2	1	2	1	2	Average
MPJPE [mm]	27.7	38.6	28.7	32.8	36.4	39.9	29.1	32.4	32.1	33.4	30.8	32.3	37.9	51.0	31.7
PCP	100	92.9	99.3	99.4	91.4	96.1	99.6	99.8	98.6	99.6	98.1	99.8	95.1	95.8	99.3
PCK@50mm	96.2	75.8	93.9	88.5	80.7	70.9	93.2	85.5	87.8	87.7	91.2	87.5	77.4	57.2	90.6
PCK@100mm	99.9	99.4	99.5	99.6	99.0	98.9	99.5	99.8	99.2	99.5	98.6	99.8	98.5	95.2	99.4
MPJPE [mm] (w/o RoM)	25.5	36.9	27.2	30.2	35.4	37.3	27.0	30.5	30.5	31.0	27.8	30.4	37.1	50.1	29.5
PCP (w/o RoM)	100	93.6	99.5	99.5	92.4	96.8	99.5	99.8	98.7	99.5	98.1	99.8	96.5	96.1	99.2
PCK@50mm (w/o RoM)	97.9	80.0	95.1	91.1	82.4	76.7	95.8	90.3	89.7	91.7	94.1	92.2	78.7	59.7	92.4
PCK@100mm (w/o RoM)	99.9	99.5	99.6	99.6	98.9	99.3	99.5	99.8	99.2	99.5	98.6	99.8	98.4	95.4	99.4

Table 3: Evaluation using the Studio dataset[1]. Ground truth is measured using an optical motion capture system. w/o RoM denotes results obtained without considering the range of motion.

5. Conclusion

The conclusions obtained from this study are following.

1. A method to realize multi-person motion capture is proposed using multiple video cameras by predicting accurate 3D pose and bounding box. This method works even in a wide field using cameras that have different fields of view placed at the same viewpoint.
2. By considering the link length, range of motion, and spatiotemporal continuity of the motion, accurate and smooth motion data is obtainable.
3. The proposed method achieves 31.7mm in mean per joint position error and 99.3% in percentage of correct parts under five people moving dynamically, while satisfying the range of motion.
4. By proposed method, all players' detailed motion who play a futsal game is acquired only from a few cameras.

Our approach still has limitations. Individual pose estimation is conducted with the bounding box in the proposed method. However, when two subjects are extremely close

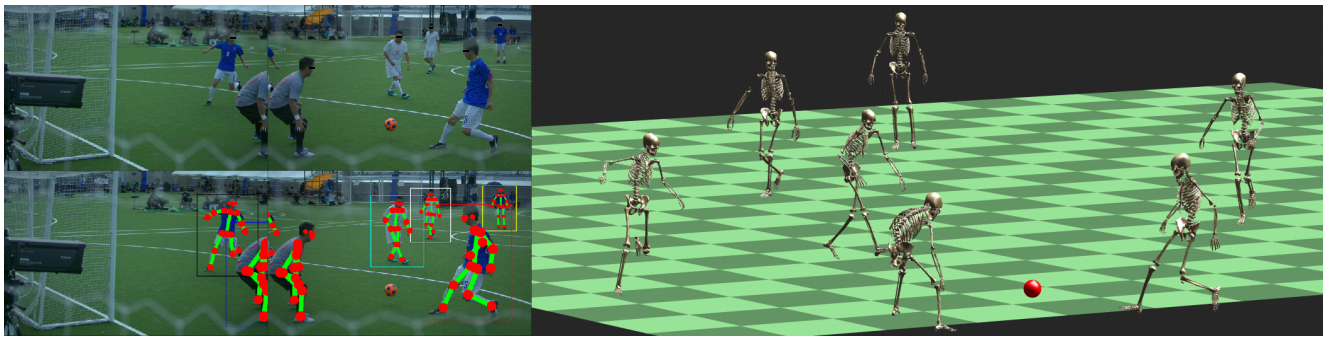


Figure 8: Qualitative result on futsal field

mutually as when hugging, the pose estimator cannot compute the part confidence maps (PCM) of the intended subject from every camera, in which leads to failure. Furthermore, when the subject moves completely out of sight of two cameras, such as when the subject is wholly occluded, or the subject comes too close to the camera, reconstruction cannot be done. However, we hope this work will guide future realization of multi-person markerless motion capture in more challenging scenes such as soccer games.

Appendix

A. Camera calibration in wide field

A 3×4 matrix M_i for projecting an arbitrary 3D point onto an image plane of camera i is expressed as:

$$M_i \equiv K_i \left[R_i | \mathbf{t}_i\right] \tag{17}$$

where K_i is an internal parameter, and R_i and $\mathbf{t}_i$ are external parameters respectively representing the attitude and position of the camera. Because the distortion parameter is calculable together with the internal parameter, it is assumed below that the internal and distortion parameter are calculated using the chess pattern [44], and the input image is compensated in advance.

External parameters are acquired using Structure from Motion (SfM) [20, 28] by the following steps.

1. Set cameras at each viewpoint; roughly estimate external parameters of each camera.
2. Move a colored sphere to cover the measurement area. Then detect the center of the sphere from multiple synchronized cameras. Triangulate them in 3D while removing the outlier using RANSAC.
3. By bundle adjustment, optimize the attitude and position of the cameras and the 3D positions of the sphere [37]. Using this method, we treat the rotation matrix, translation vector, and focal length as variables, then apply Ceres Solver for bundle adjustment [7].
4. Transform the absolute position, attitude, and scale to world coordinates while maintaining the relative relation between the cameras.

Camera calibration is performed using the process described above. A projection matrix M_i is obtained from each camera. A pixel position where a point X is projected on the image plane of the camera i is expressed as shown below.

$$\mu_i(X) = \begin{pmatrix} \left[M_i X\right]_x / \left[M_i X\right]_z \\ \left[M_i X\right]_y / \left[M_i X\right]_z \end{pmatrix} \tag{18}$$

B. Skeletal model and joint position initialization

To compute IK, the skeletal model's adjacent joints must be connected by a constant-length link. The link length must be calculated according to the human subject. Furthermore, because IK is based on iterative computation, it is reasonable to calculate the skeletal model's initial joint position before computation. In the proposed method, using multi-camera images, the pixel locations of the keypoint detected from HRNet at initial frame are reconstructed in 3D. The length parameters and initial joint position are calculated from them simultaneously.

Only the initial bounding box position is given manually. Also, because the keypoints are fewer than the number of joints, at the initial frame, restrictions such as the unbent spine and the not-raised scapula are added to subjects. In addition, the parameters are restricted so that the left and right lengths are symmetrical.

Acknowledgements

This work was made using sDIMS, a programming library for multi-body kinematics and dynamics with the human musculo-skeletal model developed in the University of Tokyo. The authors acknowledge the supports by Ayaka Yamada, Hiroki Obara, Tomoyuki Horikawa and the other students in the futsal motion capture experiment. We also thank the anonymous participants in the studio motion capture experiment. This work was conducted in the research funded by JSPS Grants-in-Aid for Scientific Research (A) JP17H00766 (2017-2019) and by NTT DOCOMO, Inc.

References

[1] Motion Analysis Corporation. http://www.motionanalysis.com.

[2] Noitom Ltd. http://neuronmocap.com/.

[3] RADiCAL. http://getrad.co/.

[4] The Captury. http://www.thecaptury.com.

[5] VICON Corporation. http://www.vicon.com/.

[6] Xsens Technologies. http://www.xsens.com/.

[7] S. Agarwal, K. Mierle, and Others. Ceres Solver. http://ceres-solver.org.

[8] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] K. Ayusawa and Y. Nakamura. Fast inverse kinematics algorithm for large DOF system with decomposed gradient computation based on recursive formulation of equilibrium. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[10] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D Pictorial Structures for Multiple Human Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[11] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D Pictorial Structures Revisited: Multiple Human Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1929–1942, Oct 2016.

[12] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *European Conference on Computer Vision (ECCV)*, 2016.

[13] L. Bridgeman, M. Volino, J.-Y. Guillemaut, and A. Hilton. Multi-Person 3D Pose Estimation and Tracking in Sports. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2019.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel, and S. Thrun. Performance Capture from Sparse Multi-view Video. *ACM Transactions on Graphics*, 27(3):98:1–98:10, Aug 2008.

[17] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou. Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] S. Ershadi-Nasab, E. Noury, S. Kasaei, and E. Sanaei. Multiple human 3D pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12):15573–15601, Jun 2018.

[19] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. LiveCap: Real-time Human Performance Capture from Monocular Video. *ACM Transactions on Graphics*, 38(2):14:1–14:17, 2019.

[20] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[21] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[22] H. Joo, t. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end Recovery of Human Shape and Pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] S. Kreiss, L. Bertoni, and A. Alahi. PifPaf: Composite Fields for Human Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[25] T-Y Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. *The Computing Research Repository*, abs/1405.0312, 2014.

[26] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[27] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics*, 36(4), July 2017.

[28] J. R. Mitchelson and A. Hilton. Wand-based Multiple Camera Studio Calibration. In *Centre for Vision, Speech and Signal Processing (CVSSP)*, 2003.

[29] F. Moreno-Noguer. 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] A. Murai, K. Kurosaki, K. Yamane, and Y. Nakamura. Musculoskeletal-see-through mirror: Computational modeling and algorithm for whole-body muscle activity visualization in real time. *Progress in Biophysics and Molecular Biology*, 103(2):310–317, 2010. Special Issue on Biomechanical Modelling of Soft Tissue Motion.

[31] T. Ohashi, Y. Ikegami, K. Yamamoto, W. Takano, and Y. Nakamura. Video Motion Capture from the Part Confidence Maps of Multi-Camera Images by Spatiotemporal Filtering Using the Human Skeletal Model. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[32] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time Human Pose Recognition in Parts from Single Depth Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[33] C. Stoll, N. Hasler, J. Gall, H. Seidel, and C. Theobalt. Fast Articulated Motion Tracking Using a Sums of Gaussians Body Model. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[34] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[35] W. Takano and Y. Nakamura. Synthesis of Whole Body Motion with Pose-Constraints from Stochastic Model. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[36] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3D Full Human Bodies Using Kinects. *IEEE Transactions on Visualization and Computer Graphics*, 18(4):643–650, April 2012.

[37] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle Adjustment – A Modern Synthesis. *Vision Algorithms: Theory and Practice*, pages 298–372, 2000.

[38] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] B. Xiao, H. Wu, and Y. Wei. Simple Baselines for Human Pose Estimation and Tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

[41] K. Yamane, Y. Fujita, and Y. Nakamura. Estimation of physically and physiologically valid somatosensory information. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2005.

[42] K. Yonemoto, S. Ishigami, and T. Kondo. Measurement Method for Range of Joint Motion (Japanese). *The Japanese Journal of Rehabilitation Medicine*, 32(4):207–217, 1995.

[43] S.-H. Zhang, R. Li, X. Dong, P. L Rosin, Z. Cai, H. Xi, D. Yang, H.-Z. Huang, and S.-M. Hu. Pose2Seg: Detection Free Human Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] Z. Zhang. A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, Nov. 2000.

[45] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.