Human Pose as Calibration Pattern; 3D Human Pose Estimation with Multiple Unsynchronized and Uncalibrated Cameras

Kosuke Takahashi Dan Mikami Mariko Isogawa Hideaki Kimata
NTT Media Intelligence Laboratories
NIPPON TELEGRAPH AND TELEPHONE COORPORATION, Japan
{kosuke.takahashi.rd, dan.mikami.vp, mariko.isogawa.kt, hideaki.kimata.yu}@hco.ntt.co.jp

Abstract

This paper proposes a novel algorithm of estimating 3D human pose from multi-view videos captured by unsynchronized and uncalibrated cameras. In a such configuration, the conventional vision-based approaches utilize detected 2D features of common 3D points for synchronization and camera pose estimation, however, they sometimes suffer from difficulties of feature correspondences in case of wide baselines. For such cases, the proposed method focuses on that the projections of human joints can be associated each other robustly even in wide baseline videos and utilizes them as the common reference points. To utilize the projections of joint as the corresponding points, they should be detected in the images, however, these 2D joint sometimes include detection errors which make the estimation unstable. For dealing with such errors, the proposed method introduces two ideas. The first idea is to relax the reprojection errors for avoiding optimizing to noised observations. The second idea is to introduce an geometric constraint on the prior knowledge that the reference points consists of human joints. We demonstrate the performance of the proposed algorithm of synchronization and pose estimation with qualitative and quantitative evaluations using synthesized and real data.

1. Introduction

Measuring 3D human pose is important for analyzing the mechanics of the human body in various research fields, such as biomechanics, sports science and so on. In general, some additional devices, *e.g*. optical markers [1] and inertial sensors [2], are introduced for measuring 3D human pose. While these approaches have advantages in terms of high estimation quality, *i.e*. precision and robustness, it is sometimes difficult to utilize them in some practical scenarios, such as monitoring people in daily life or evaluating the performance of each player in a sports game, due to inconveniences of installing the devices.

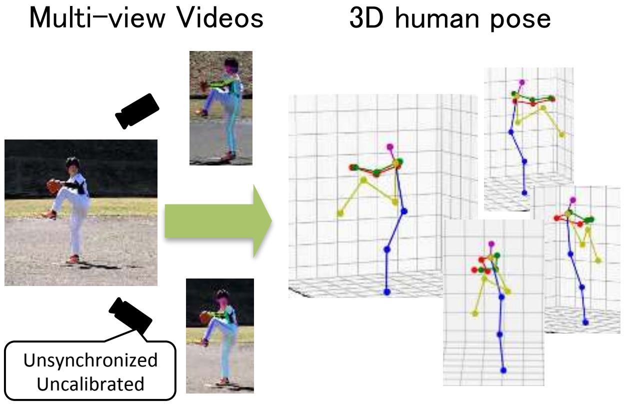

Figure 1. Our target. 3D human pose estimation with unsynchronized and uncalibrated cameras with wide baselines.

To estimate 3D human pose in such cases, vision-based motion capture techniques have been studied in the field of computer vision [9]. Basically, they utilize multi-view cameras or depth sensors and assume that they are synchronized and calibrated beforehand. Such synchronization and calibration are troublesome to establish and maintain; typically, the cameras are connected by wires and capture the same reference object. Some 2D local features based synchronization and calibration methods have also been proposed for easy-to-use multi-view imaging systems in case for which the preparation cannot be done. However, they sometimes suffer from difficulties of feature correspondences in case for which the multiple cameras are scattered with wide baselines, which the erroneous correspondences affect the stability and precision of estimation severely.

This paper addresses the problem of 3D human pose estimation from multi-view videos captured by unsynchronized and uncalibrated cameras with wide baselines. The key feature of this paper is its focus on using the projections of human joints to derive robust point associations for use as common reference points. To detect the projections of human joints some 2D form of pose detector is needed [7,24,25], however, 2D joint positions sometime include detection errors which make the estimation unstable. To deal with such errors, the proposed method introduces two ideas. The first idea is to relax the reprojection errors to avoid the optimization of noisy observations. The second idea is to introduce a geometric constraint based on the a priori knowledge that the reference points are actually human joints.

The key contribution of this paper is to propose a novel algorithm for 2D human joint based multi-camera synchronization, camera pose estimation and 3D human pose estimation. This algorithm enables us to obtain 3D human pose easily and stably even in a practical and challenging scenes, such as sports games.

The reminder of this paper is organized as follow. Section 2 reviews related works in terms of synchronization, extrinsic calibration and human pose estimation. Section 3 introduces our proposed algorithm using the detected 2D joint positions as the corresponding points in multi-view images. Section 4 reports the performance evaluations and Section 5 provides discussions on the proposed method. Section 6 concludes this paper.

2. Related Works

This section introduces the related works of our research in terms of (1) camera synchronization, (2) extrinsic camera calibration, and (3) human pose estimation.

Camera Synchronization Multiple camera synchronization significantly impacts the estimation precision of multi-view applications, such as 3D reconstruction. In general, the cameras are wired and receive a trigger signal from an external sensor telling the camera when to acquire an image. However, these wired connections can be troublesome to establish when the cameras are widely scattered as happens when capturing a sports games.

Audio-based approaches [9,18] estimate the time shift of multi-view videos in a software post-processing step, however, the significant difference in camera position degrades the estimation precision due to the delay in sound arrival.

Some image-based methods [3,22] are able to synchronize the cameras even in such cases. Cenek *et al.* [3] estimates the time shift by using epipolar geometry on the corresponding points in the multi-view images. Tamaki *et al.* [22] detected the same table tennis ball in sequential frames and utilized them to establish point correspondences for computing epipolar geometry. Given the scale of the capture environment envisaged, our method is also based on epipolar geometry and so uses detected 2D joint positions as the corresponding points.

Extrinsic Camera Calibration Extrinsic camera calibration is an essential technique for 3D analysis and understanding from multi-view videos and various proposals have been made for various camera settings. Most proposals utilize detected 2D features, such as chess board corners or local features, *e.g.* SIFT [13], as the corresponding points. These approaches have difficulty in establishing reliable feature correspondence if the multiple cameras are scattered with wide baselines, as erroneous correspondences degrade the stability and precision of estimation severely.

For such cases, some studies utilize a priori knowledge of the scene. Huang *et al.* [11] use the trajectories of pedestrians in calibrating multiple fixed cameras based on the assumption that the cameras can capture the same pedestrians for a long time. Namdar *et al.* [10] assume that the cameras capture a sport scene in a stadium and calibrate them by introducing vanishing points computed from the lines on the sports field.

In addition, some studies [6, 16, 20] propose calibration algorithm that utilizes a priori knowledge that that the scenes contain humans. The silhouette-based approaches [5,19] establish the correspondences between special points on the silhouette boundaries, called *frontier points* [8], across the multiple views. These points are the projections of 3D points tangent to the epipolar plane. The epipolar geometry can be recovered from the correspondences of the frontier points.

Puwein et al. [16] proposed using detected 2D human joints in multi-view images as common reference points and using these points to compute the extrinsic parameters. Our method is inspired by [16]. In [16], the error function consists of reprojection error, a kinematic structure term, a smooth motion term and so on, is minimized in the bundle adjustment manner. Our work, on the other hand, introduces a relaxed reprojection error for robust estimation in the face of very noisy data; it also solves the synchronization problem.

2D Human Pose Estimation Conventional studies of 2D human pose estimation problem fall into two basic groups: pictral structure approach [4, 15], in which spatial correlations between each part are expressed as a tree-structured graphical model with kinematic priors that couple connected limbs, and hierarchical model approach [21, 23], which represents the relationships between parts at different scales and sizes in a hierarchical tree structure.

Given the rapid improvement in neural network techniques, a lot of neural network based 2D pose detectors have been proposed [7,24,25]. Toshev *et al.* [24] solve 2D human pose as a regression problem by introducing the AlexNet architecture, which was originally used for object recognition. Wei *et al.* [25] achieve high precise pose estimation by introducing CNN to the Pose Machine [17]. Cao*et al.* [7]

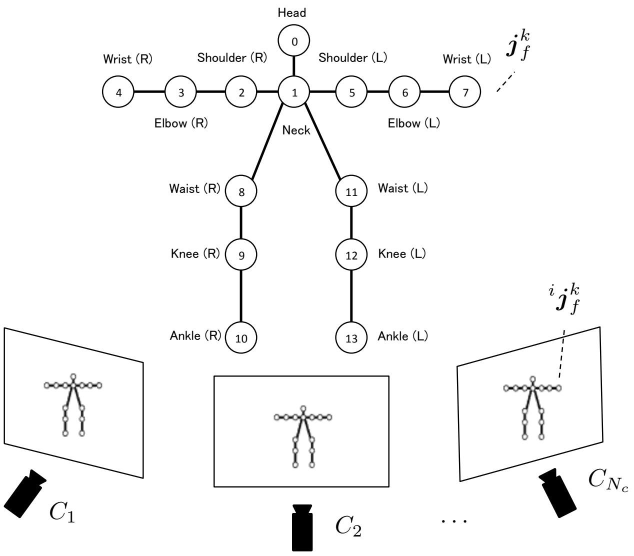

Figure 2. Configuration of the capture system. These cameras capture a human modeled as a set of articulated 3D joints $\boldsymbol{j}_f^k$.

consider the connectivity of each joint by introducing a part affinity field to the work of [25]; they achieve robust estimation of multi-person pose in real time.

3. Proposed Method

This section describes our proposed method for estimating 3D human pose with unsynchronized and uncalibrated multiple cameras with wide baselines.

3.1. Problem Formulation

This paper assumes that a human body is captured by multiple unsynchronized and uncalibrated cameras. As illustrated in Figure 2, the human body is modeled as a set of articulated 3D joints. The 3D position of the k th joint and its 2D projection onto the image plane of the ith camera, C_i, in frame f are represented as $J = \{\boldsymbol{j}_f^k\}, k \in [1, \cdots, N_J]$ and ${}^{i}\boldsymbol{j}_f^k, i \in [1, \cdots, N_c, N_c \geq 2]$ respectively.

Let $P = \{R_i, \boldsymbol{t}_i\}$ denote the rotation matrix and translation vector, that is the extrinsic parameters of ith camera C_i; they satisfy,

$$\boldsymbol{p}^{C_i} = R_i \boldsymbol{p}^W + \boldsymbol{t}_i \tag{1}$$

where $\boldsymbol{p}^{C_i}$ and $\boldsymbol{p}^W$ denote the coordinates of 3D point $\boldsymbol{p}$ in the C_i coordinate system and the world coordinate system respectively. In this paper, C_1 is the base camera and its coordinate system is used as the world coordinate system.

Let $D = \{d_i\}$ denote the temporal difference in frame scale compared with the base camera C_1. This d_i satisfies the following equation,

$$f_t^0 = f_t^i + d_i \tag{2}$$

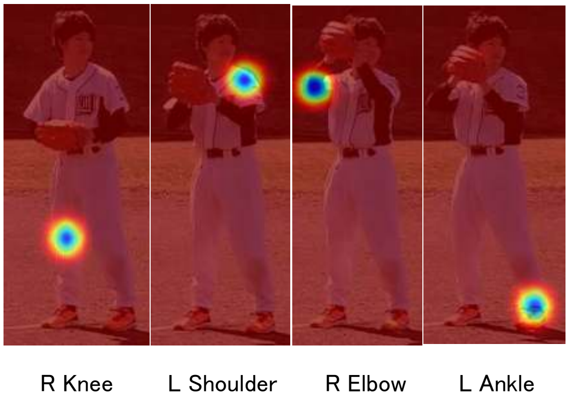

Figure 3. Confidence map of each joint. Blue area represents a high-confidence area and red area represents a low-confidence area.

where f_t^i denotes a $t \in [1, \cdots, N_t]$ th frame of a video captured by C_i. Hereafter, f_t^0 is written as f_t for simplicity.

The goal of this research is to estimate the 3D positions of human joint $\boldsymbol{j}_f^k$, the extrinsic camera parameters R_i and $\boldsymbol{t}_i$, and temporal differences d_i. This paper assumes that a single human appears in the captured video, however, the proposed method can be extended to cover multiple people. This extension is discussed in Section 5.

For estimating these parameters, the proposed method regards as the human model as a reference object and takes a bundle adjustment approach by utilizing their projections ${}^{i}\boldsymbol{j}_f^k$ as points for which correspondence is to be found. The proposed method defines the following objective function,

$$\underset{P,J,L,D}{\arg\min}\, E(P, J, L, D) \tag{3}$$

where L denotes the separation of each joint pair, introduced in Section 3.1.2, and minimizes Eq.(3) over parameters P, J, L and D.

This objective function consists of two major error terms as follows,

$$E(P, J, L, D) = E_{rep}(P, J, D) + E_{mdoel}(J, L, D) \tag{4}$$

where $E_{rep}(P, J, D)$ and $E_{model}(J, L, D)$ represent the error terms of the reprojection error and the human model, respectively. Following sections detail these error terms.

3.1.1 Relaxed Reprojection Error

The conventional 2D features for camera synchronization or calibration, such as chess corners, local features and so on, are detected with sub-pixel precision. On the other hand, the proposed method utilizes the 2D joint positions detected by a 2D pose estimation algorithms [7, 24, 25] as 2D features and most of these positions include detection errors of a few pixels. These detection errors severely impact the performance of the conventional bundle adjustment approach,

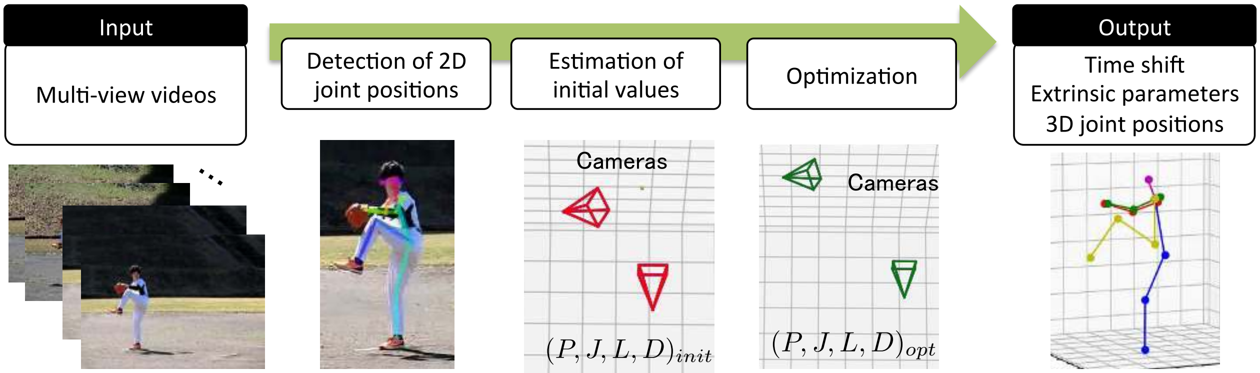

Figure 4. Outline of the proposed method. Firstly the 2D joint positions are detected from the multi-view videos, and the initial values of each parameter are estimated in a SfM manner with assumed time shift d_i. Next, the initial values are optimized in terms of relaxed reprojection error and constraints on human joint. Finally, the parameters yielding the smallest values of Eq. (4) with d_i are selected as the output of the proposed method.

which attempts to minimize the reprojection errors. Here, the proposed method avoids the problem of detection errors by relaxing the reprojection errors.

Most conventional 2D pose estimation techniques such as [7, 24, 25] estimate a confidence map for each joint and define the 2D joint position as the peak of the map as illustrated in Figure 3. Following this idea, the proposed method uses the confidence map to relax the reprojection error; that is, the influence of the reprojection error is weakened when the reprojected point is in an area of high-confidence and enhanced in when the reprojected point is in an area of low-confidence. The proposed method assumes that the high-confidence area in the confidence map follows a normal distribution and defines the reprojection error term as follows,

$$E_{rep}(P, J, D) = \frac{1}{N_{\,rep}} \Sigma_{t=0}^{N_t} \Sigma_{i=0}^{N_c} \Sigma_{k=0}^{N_j} g({}^{k}\boldsymbol{j}_{f_t}^{i}, {}^{k}\tilde{\boldsymbol{j}}_{f_t}^{i}), \quad (5)$$

where $N_{rep} = N_t \times N_c \times N_j$ and ${}^{k}\tilde{\boldsymbol{j}}_{f_t}^{i}$ denote the reprojection of ${}^{k}\boldsymbol{j}_{f_t}^{i}$ computed from P, J and D, and

$$g(\boldsymbol{x}, \boldsymbol{x}') = (n(0) - n(e_{rep}(\boldsymbol{x}, \boldsymbol{x}')))e_{rep}(\boldsymbol{x}, \boldsymbol{x}'), \quad (6)$$

$$e_{rep}(\boldsymbol{x}, \boldsymbol{x}) = ||\boldsymbol{x} - \boldsymbol{x}'||. \quad (7)$$

$n(x)$ denotes the probability density function of normal distribution $N(\mu_p, \sigma_p^2)$ and $||\boldsymbol{x}||$ denotes the L^2-norm of $\boldsymbol{x}$.

3.1.2 Constraints on Human Joints

The proposed method assumes that the multi-cameras capture a human body and introduces constraints based on a priori knowledge as $E_{model}(J, L, D)$. Error term $E_{model}(J, L, D)$ has two terms as follows,

$$E_{model}(J, L, D) = E_{length}(J, L, D) + E_{motion}(J, D). \quad (8)$$

The following sections describe these error terms in detail.

Constraint on Length of a Joint Pair The pair of the k th joint and the k' joint is denoted as $\langle k, k' \rangle$ in Figure 2. The pairs of $\langle 2, 3 \rangle$ and $\langle 8, 9 \rangle$ can be recognized as the humerus and femur, respectively, and the length between the 3D joints on each bone are taken to be constant over time. Here, the proposed method assumes the joint pairs $P = \{\langle 0, 1 \rangle, \langle 1, 2 \rangle, \langle 1, 5 \rangle, \langle 2, 3 \rangle, \langle 3, 4 \rangle, \langle 5, 6 \rangle, \langle 6, 7 \rangle, \langle 8, 9 \rangle, \langle 8, 11 \rangle, \langle 9, 10 \rangle, \langle 11, 12 \rangle, \langle 12, 13 \rangle\}$ has consistent length and introduce the error term $E_{length}(J, L, D)$ as follows,

$$E_{length}(J, L, D) = \Sigma_t^{N_t} \Sigma_P |||\boldsymbol{j}_t^k - \boldsymbol{j}_t^{k'}|| - l(\langle k, k' \rangle)|, \quad (9)$$

where $l(\langle k, k' \rangle)$ represents the distance between joint pair $\langle k, k' \rangle$ and $L = \{l(\langle k, k' \rangle)\}$.

Constraint on Smooth Motion of Each Joint The proposed method introduces a constraint on the smooth motion of a joint based on the observation that the 3D positions the joints do change drastically in sequential frames. The proposed method assumes that the local motion of each joint can be modeled as the linear motion created by uniform acceleration and introduces the following error term,

$$E_{motion}(J, D) = \frac{1}{N_t \times N_j} u(\boldsymbol{j}_t^k). \quad (10)$$

$u(\boldsymbol{j}_t^k)$ represents the third order differential value of $\boldsymbol{j}_t^k$. The minimization of $u(\boldsymbol{j}_t^k)$ forces the second order differential value of $\boldsymbol{j}_t^k$, that is the acceleration, to be consistent in sequential frames.

3.2. Algorithm

Figure 4 illustrates the processing flow of the proposed method. First, it detects 2D joint positions from the input multi-view videos using a 2D pose detector such as

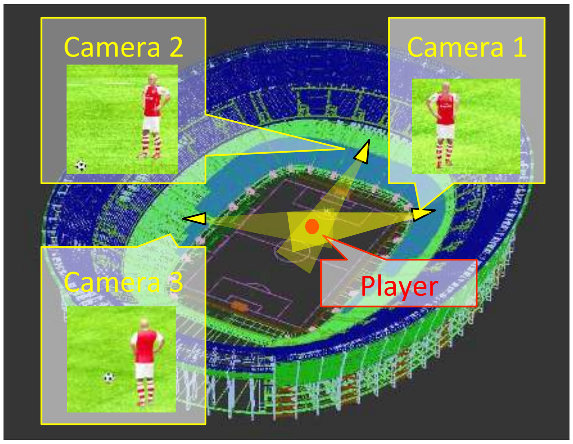

Figure 5. Configuration of evaluation with synthesized data. Three cameras are set around a field with wide baselines.

[7,24,25]. Since the 2D pose detector output includes detection errors and joint detection sometimes fails due to self-occlusion, the proposed method applies a median filter after applying a cubic spline interpolation method to the output data. Next, select two cameras and the initial values of each parameter are estimated by the standard SfM approach using assumed time shift d_i; that is, it estimates the essential matrix for selected cameras, decomposes it into the extrinsic parameters, and estimates 3D joint positions through triangulation. The extrinsic camera parameters of the other cameras are estimated by solving PnP problems [12]. Then, Eq.(4) is computed using the initial parameters and minimized by the Levenberg-Marquardt algorithm over parameters P, J, and L. Finally, the parameters yielding the smallest value of Eq. (4) with d_i are selected as the optimized parameters.

4. Evaluations

This section describes the performance evaluations of the proposed method with synthesized data and real data.

4.1. Evaluations with Synthesized Data

4.1.1 Experimental Environment

Figure 5 illustrates the evaluation setup used with synthesized data. The three unsynchronized cameras are set around a large field with baselines of about $50 \sim 100$m. These cameras capture 1920×1080 resolution videos with 60 frame rate. Their focal length and optical center, that is intrinsic parameters, are set to 16000 and $(960, 540)$ respectively. The ground truth of 3D joint is synthesized from the motion capture data. The input data, the projections of the 3D joint positions, is perturbed by the addition of zero-mean Gaussian noise whose standard deviation $\sigma (0 \leq \sigma \leq 8)$. The input data also includes 10% detection failures.

To demonstrate the performance of the proposed algorithm, the following two conventional methods are evaluated with same input data,

Method1: Initial values As described in Section 3.2, each parameter can be linearly estimated in a conventional structure-from-motion manner. As to Method1, the evaluation function for selecting appropriate time shift is defined as the reprojection error, that is the parameters with smallest reprojection error are the output of Method1.

Method2: Bundle Adjustment Method2 uses the bundle adjustment approach to estimate the parameters, each parameter is optimized by minimizing the reprojection errors. Here, Method2 uses the Levenberg-Marquardt algorithm for optimization. Method2 also utilizes the reprojection error as the evaluation function for selecting the appropriate time shift.

In this evaluation, each parameter is evaluated with following error functions. The error of time shift E_f is defined as the average of absolute error (millisecond time scale) as follows,

$$E_f = \frac{1}{N_c} \Sigma_{i=1}^{N_c} |f_i - f_{ig}|, \tag{11}$$

where the parameter with subscript g represents the ground truth. The error of rotation matrix E_R is defined as the Riemannian distance [14].

$$E_R = \frac{1}{N_i \sqrt{2}} \Sigma_{i=1}^{N_c} ||\text{Log}(R_i^\top R_{ig})||_F, \tag{12}$$

$$\text{Log} R' = \begin{cases} 0 & (\phi = 0), \\ \frac{\phi}{2 \sin \phi}(R' - R'^\top) & (\phi \neq 0). \end{cases} \tag{13}$$

where $\phi = \cos^{-1}(\frac{\text{tr}R' - 1}{2})$ and $|| \cdot ||_F$ denotes Frobenius norm. The error of translation vector E_t is defined as

$$E_t = \frac{1}{N_c} \Sigma_{i=1}^{N_c} ||s_i^t \boldsymbol{t}_i - \boldsymbol{t}_{ig}||, \tag{14}$$

where s_i^t denotes a scale parameter computed by $s = ||\boldsymbol{t}_i|| / ||\boldsymbol{t}_{ig}||$. The error of 3D joint, E_j, is defined as,

$$E_j = \frac{1}{N_j \times N_t} \Sigma_{k=1}^{N_j} \Sigma_{f=1}^{N_f} ||s^j \boldsymbol{j}_f^k - \boldsymbol{j}_f^k|| \tag{15}$$

where s^j represents a scale parameter that makes each estimated joint pair length match its ground truth.

4.1.2 Results

Figure 6 plots the average errors of synchronization, extrinsic parameters, and 3D joint positions for 10 trials at each noise level. From these results, we can see that all methods

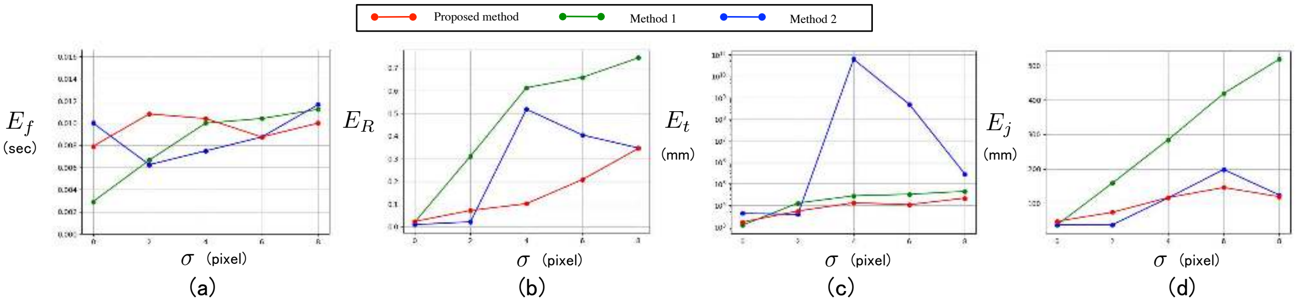

Figure 6. Estimation error of (a) time shift, (b) rotation matrix, (c) translation vector, and (d) 3D human positions.

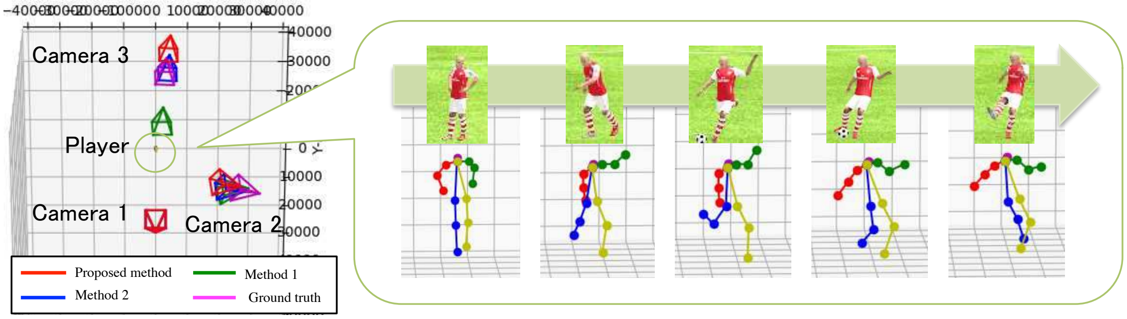

Figure 7. Visualization of estimated camera positions and 3D joint positions in $\sigma = 5$ case. Left; Camera positions estimated by the proposed method (red), Method1 (green), Method2 (blue) and its ground truth (magenta). The positions of C_1 estimated by each method are set to $(0, 0, 0)^\top$. Right; 3D positions of each joint. The upper figures show the ground truth data with a CG human model in each frame and the lower figures show the 3D joints estimated by the proposed method in each corresponding frame.

estimate the time shift with error of $0.006 \sim 0.012$ seconds. As to the extrinsic parameters, the proposed method offers robust estimation even if large detection error is assumed while the conventional methods suffer degraded performace. Especially, Method2 significantly degrades in $\sigma > 1$ cases. We consider that the reason is that the Method2, which minimizes the reprojection errors strictly, is significantly affected by the noise and detection failures. The proposed method estimates 3D joint positions robustly while the Method1 degrades with noisy data. Method2 also estimates 3D joint positions with comparable precisions to the proposed method in spite of its degraded extrinsic parameters. This is considered that the adjusting the scale and initial position in case of evaluating the 3D positions absorbs this degradation.

Figure 7 renders one example of the estimated camera positions and 3D joint positions in 3D space with $\sigma = 5$. This figure shows that the proposed method estimates reasonable camera positions and 3D joints.

From the above, we can conclude that the proposed method is more robust than the conventional methods even if the input data includes significant noise, especially in terms of the extrinsic parameters.

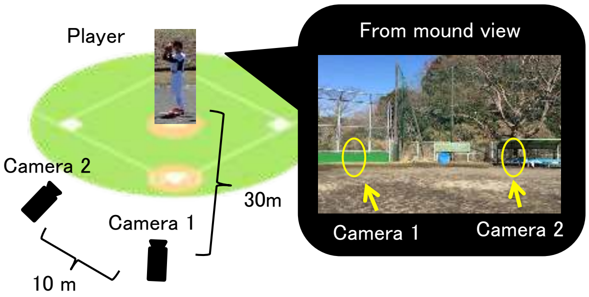

Figure 8. Configuration of evaluation with real data. Three cameras are set around a field with wide baselines.

4.2. Evaluations with Real Data

This section demonstrates that the proposed method works with real data in a practical scenario.

4.2.1 Experimental Environment

Figure 8 shows the configuration of the evaluation that used real data. The two cameras (CASIO EX100) with 640 × 480 resolution and 120 fps were set with a wide baseline. Camera 0 and Camera 1 had focal lengths of 200mm and 165mm, respectively. The input video consisted of 1000 frames, that is about 8.3 seconds. These cameras captured one player throwing a ball.

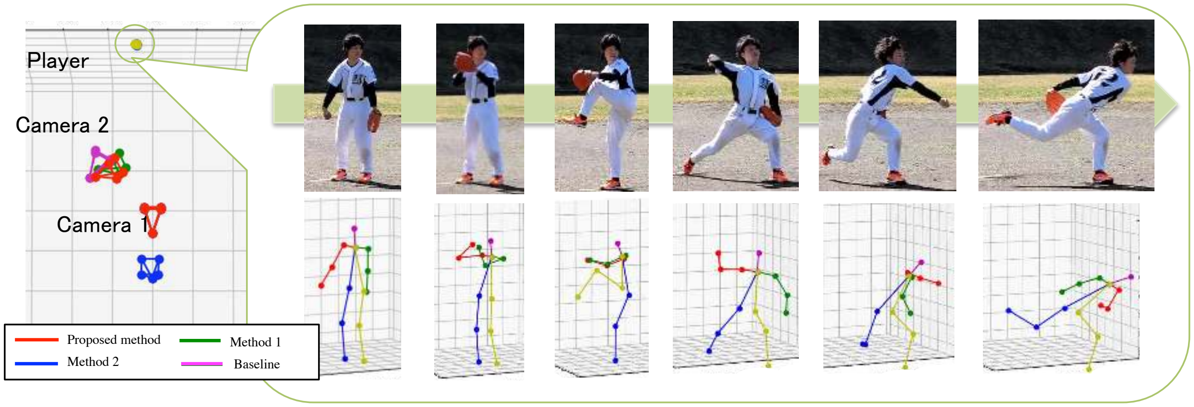

Figure 9. Visualization of estimated camera positions and 3D joint positions with the real data. Left; Camera positions estimated by the proposed method (red), Method1 (green), Method2 (blue) and baseline (magenta). The positions of C_1 estimated by each method are set to $(0,0,0)^\top$. Right; 3D positions of each joint. The upper figures show input data captured by camera 2 in each frame and the lower figures show the 3D joints estimated by the proposed method in each corresponding frame.

Table 1. Evaluation of extrinsic parameters.

	E_R	E_t (mm)
Method1	0.4162	3491
Method2	0.5843	10973
Proposed method	0.3683	5355

In the 2D pose estimation step, Cao *et al*. [7]'s method is utilized. This evaluation used, in addition to Method1 and Method2 introduced in Section 4.1, Zhang's method [26] as benchmarks.

4.2.2 Results

Table 1 reports the estimation error of extrinsic parameters between each method and [26]. Figure 9 visualizes the estimated camera positions and 3D joint positions in 3D space. In Figure 9, while the camera positions estimated by Method1, Zhang's method and the proposed method are almost same, that of Method2 diverged significantly. The reason is that the detected 2D poses include severe noise and Method2, which minimizes the reprojection errors strictly, is optimized to the noisy data same as in the evaluations with synthesized data, whereas the proposed method avoids the problem by relaxing the reprojection errors.

From these results, we can see that the proposed method works robustly with severely noise-degraded data in practical situations.

5. Discussion

5.1. Precision of 2D Human Pose Detector

The 2D pose detector is utilized in the first step of the proposed method and it has significant effects on the estimation precision. Here we investigate the performance of 2D pose detector [7] utilized in this paper.

Table 2. Evaluation of 2D pose detector (pixel).

Ave	Std	Min	Max
8.144	6.188	0.0058	48.492

Table shows the average, standard deviation, smallest value and biggest value of estimation error, that is euclidean norm of 2D human pose detected by [7] and its ground truth, in 700 frames with 1920 × 1080 resolutions. In the evaluations in Section 4, the μ_p and σ_p in the relaxed reprojection error are set based on these results.

5.2. Multi-player Cases

As introduced in Section 3.1, our algorithm assume that there is a single player in the shared field-of-view of multiple cameras, however, it can be extend to multi-player cases. By considering the multi-players, it is considered that the estimation precision by the proposed method is improved because the number of constraints increase and the 2D joint positions, which are recognized as the corresponding points, cover more wide area in image planes of each camera. To deal with multi-player cases, the person identification problem and occlusion handling are to be solved in addition. This extension is included in our future works.

6. Conclusion

This paper proposed a novel 3D human joint position estimation algorithm for unsynchronized and uncalibrated cameras with wide baselines. The method focuses on the major skeleton joints and the constancy of joint separation. The 2D human pose is detected from the multi-view images and joint position estimates are used in the structure-from-motion manner. The proposed method provides an objective function consisting of a relaxed reprojection error term and human joint error term in order to achieve robust estimation even if the input data is noisy; the objective term is optimized. Evaluations using synthesized data and real data showed that the proposed method works robustly with noise-corrupted data. Future works include evaluations that use marker-based motion capture techniques and extension to the multi-player cases.

References

[1] *OptiTrack*. `http://optitrack.com/`.

[2] *Xsens MVN*. `https://www.xsens.com/`.

[3] C. Albl, Z. Kukelova, A. Fitzgibbon, J. Heller, M. Smid, and T. Pajdla. On the two-view geometry of unsynchronized cameras. *arXiv preprint arXiv:1704.06843*, 2017.

[4] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1014–1021. IEEE, 2009.

[5] G. Ben-Artzi, Y. Kasten, S. Peleg, and M. Werman. Camera calibration from dynamic silhouettes using motion barcodes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4095–4103, 2016.

[6] E. Boyer. On using silhouettes for camera calibration. *Proc. Asian Conf. on Computer Vision*, pages 1–10, 2006.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, July 2017.

[8] R. Cipolla and P. Giblin. *Visual motion of curves and surfaces*. Cambridge University Press, 2000.

[9] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 224–231. IEEE, 2009.

[10] N. Homayounfar, S. Fidler, and R. Urtasun. Sports field localization via deep structured models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[11] S. Huang, X. Ying, J. Rong, Z. Shang, and H. Zha. Camera calibration from periodic motion of a pedestrian. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 2016.

[12] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81(2), 2008.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] M. Moakher. Means and averaging in the group of rotations. *SIAM J. Matrix Anal. Appl.*, 24:1–16, 2002.

[15] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 588–595. IEEE, 2013.

[16] J. Puwein, L. Ballan, R. Ziegler, and M. Pollefeys. Joint camera pose estimation and 3d human pose estimation in a multi-camera setup. In *Proc. Asian Conf. on Computer Vision*, pages 473–487. Springer, 2014.

[17] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *Proc. European Conf. on Computer Vision*, pages 33–47. Springer, 2014.

[18] P. Shrstha, M. Barbieri, and H. Weda. Synchronization of multi-camera video recordings based on audio. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 545–548. ACM, 2007.

[19] S. N. Sinha and M. Pollefeys. Camera network calibration and synchronization from silhouettes in archived video. *International Journal of Computer Vision*, 87(3):266–283, 2010.

[20] S. N. Sinha and M. Pollefeys. Camera network calibration and synchronization fromsilhouettes in archived video. *International Journal of Computer Vision*, 87(3):266–283, May 2010.

[21] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *Proc. International Conf. on Computer Vision*, pages 723–730. IEEE, 2011.

[22] S. Tamaki and H. Saito. Reconstructing the 3d trajectory of a ball with unsynchronized cameras. *International Journal of Computer Science in Sport (International Association of Computer Science in Sport)*, 14(1), 2015.

[23] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *Proc. European Conf. on Computer Vision*, pages 256–269. Springer, 2012.

[24] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.

[25] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[26] Z. Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 1330–1334, 2000.