3D Pose Estimation Using Multiple Asynchronous Cameras

Takashi Morimoto, Ikuhisa Mitsugami

Abstract This paper proposes a method for estimating the 3D pose of a person using multiple asynchronous cameras. In the proposed method, a 2D pose of a person is estimated from each captured image using OpenPose. To solve the asynchrony problem, we virtually generate the synchronous pose data by interpolating the temporally neighboring poses that are from actually captured images. Then, the 3D pose is reconstructed by triangulation from the virtually synchronized 2D poses from multiple cameras. In the experiment, the effectiveness of the proposed method was confirmed by capturing a moving person using eight cameras. We also investigated the effect of frame rate changes on pose estimation accuracy.

1 Introduction

Motion capture systems that measure the 3D whole-body motions of a person have been used in various situations for a wide range of purposes. The movement of the human body contains important information for recognizing the person's interests and intentions, and for identifying physical problems. Such systems, however, usually have a restriction that a person has to wear a special suit to attach multiple markers/sensors on his/her body, which prevents him/her from doing comfort and natural behaviors.

There are also motion capture methods that do not require wearing special suits or sensors. For example, devices with depth sensors, such as Microsoft Kinect, can measure the 3D pose of a person. Deep learning based methods, such as OpenPose[1, 2, 3, 4], can measure 2D poses from RGB images, which from multiple viewpoints are further integrated for reconstructing the 3D pose [1]. This method, however,

Takashi Morimoto
Hiroshima City University, e-mail: morimoto@sys.info.hiroshima-cu.ac.jp

Ikuhisa Mitsugami
Hiroshima City University, e-mail: mitsugami@hiroshima-cu.ac.jp

requires physically synchronous cameras to capture images at the same moment from different viewpoints. Such cameras are usually expensive, and it is not easy to construct an environment using those cameras.

In this study, we thus propose a 3D pose estimation method for multiple asynchronous cameras. In this method, pseudo-synchronous 2D poses are generated from 2D poses that are obtained from actually captured images. Once we get those 2D poses that can be regarded to be synchronized, we apply the triangulation to each joint point to reconstruct 3D poses. In addition, to solve the self-occlusion problem, we apply RANSAC in the reconstruction process. Through experiments where eight consumer cameras are located around a person, we confirm the effectiveness of the proposed method. In addition, we investigate the pose reconstruction performance at different frame rates.

2 Related Work

There have been many studies on marker-less 3D pose estimation. Kinect is one of the popular devices for pose estimation; it is a device that can estimate the pose of a person with a built-in depth sensor, and is used in various researches[5, 6, 7, 8]. References [5, 6] proposed methods that uses multiple Kinects to perform 3D shape reconstruction of the capturing environment and 3D pose estimation of a person. These studies also address the issue of asynchronous imaging devices. References [7, 8] identify a part label of each pixel in a depth image by creating a large number of random tree decision trees based on simple features.

There are also methods using RGB cameras, and many methods for estimating the 3D pose from monocular images have been proposed. The approaches of using 2D-3D data for training has been around for several years [9, 10, 11, 12]. For example, Martinez et al. [9] proposed a method that trains 2D-3D pose correspondences using a deep neural network. Recently, several methods that utilize projection from 3D to 2D have been proposed [13, 14, 15]. For example, Wandt et al. [13] proposed a method for estimating 3D poses from a single image by weakly supervised learning using constraints on the correspondence between 2D and 3D poses. There are also studies that use images from multiple viewpoints at the learning stage and finally estimate the 3D pose from monocular images [16, 17, 18]. For example, Rhodin et al. [16] proposed a method for training a deep network using multi-view images and predicting 3D pose from a single image. However, accurate 3D poses are not guaranteed because these methods for obtaining 3D poses based on learning are only estimates.

In addition to methods using a monocular camera, there are also methods using multiple view geometry for estimating the 3D pose. For example, Cao et al. [1] implemented “3D Reconstruction Module” that reconstructs 3D poses by applying triangulation to multiple 2D poses obtained from multiple synchronous cameras. Such cameras, however, usually expensive, and it is not easy to construct an environment that consists of the synchronous cameras. On the other hand, the proposed

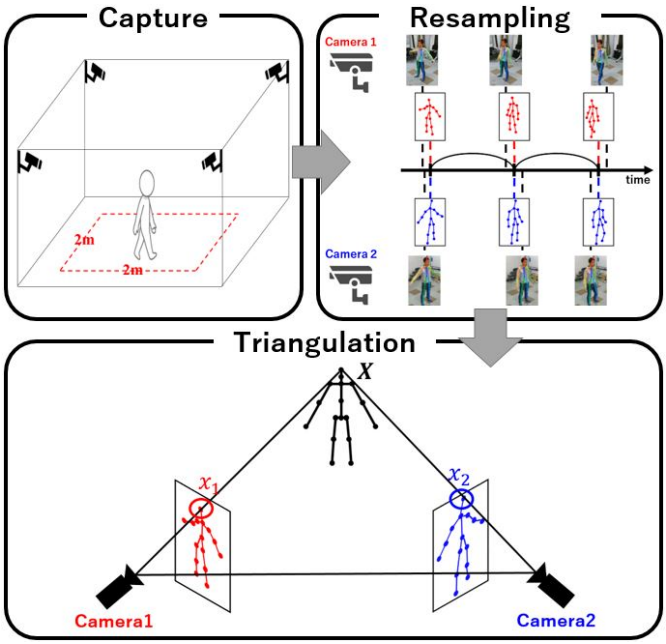

Fig. 1 Outline of the proposed method. First, 2D attitude data are obtained by applying OpenPose to RGB images captured by multiple cameras. Next, pseudo-synchronous data are generated by resampling. Finally, 3D pose data is reconstructed by triangulation.

method uses practical cameras that cannot be physically synchronized by virtually generate synchronous 2D poses from actually asynchronous ones.

3 Proposed Method

Fig. 1 shows an overview of the proposed method. We locate multiple cameras around a location where a person would stand for observing each joint point from some cameras regardless of the position and pose of the person. Each camera is calibrated in advance for eliminating lens distortion by Zhang's calibration method [19], and 2D poses are estimated from captured images to obtain 2D poses (specifically, 2D positions of joint points). Since the 3D pose estimation is based on the principle of triangulation, it is necessary to obtain the perspective projection matrices of the multiple cameras as a preparation. As the capturing timing differs depending on the cameras, resampling is performed for generating pseudo-synchronized 2D poses. Finally, 3D pose is reconstructed by apply triangulation to those pseudo-synchronized 2D poses (hereinafter, this method is called OpenPose Stereo). In this triangulation process, the pose is robustly estimated by removing joint points that are not properly estimated using the RANSAC.

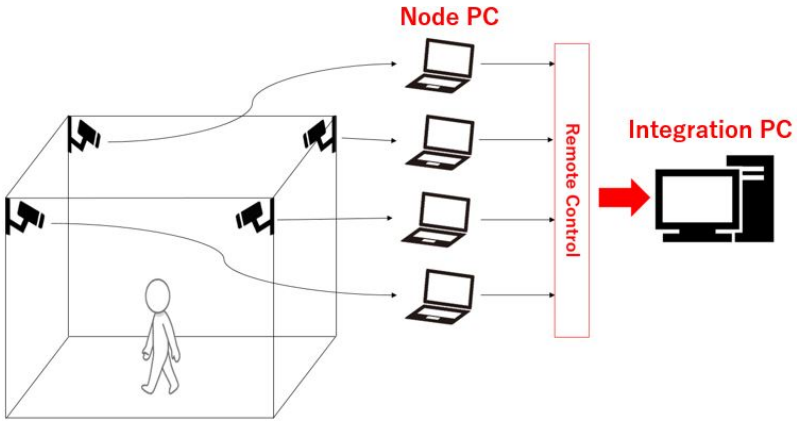

Fig. 2 System Configuration. Each camera is connected to a Node PC, and the measured data are collected on a PC for integration.

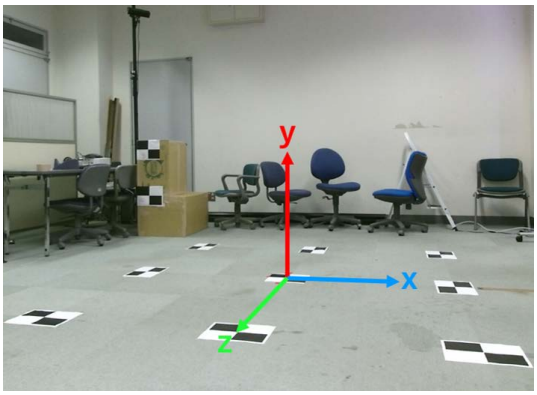

Fig. 3 Camera Calibration. Install checker patterns in the experimental environment and capture them with all cameras.

3.1 System Configuration

In this study, multiple cameras are placed around a person in order to accurately estimate the pose of a person who moves freely. Self-occlusion may occur depending on the relative positions between the camera and the person. For eliminating this risk, multiple cameras are placed at different positions and capture a person from various angles. As shown in Fig. 2, each camera is connected to a PC (we call it "Node PC") and captured images are stored in the PC with their timestamps. 2D poses are then estimated by OpenPose in each Node PC, and they are then transmitted to the Integration PC, where the 3D pose is recovered from the 2D poses sent from the Node PCs.

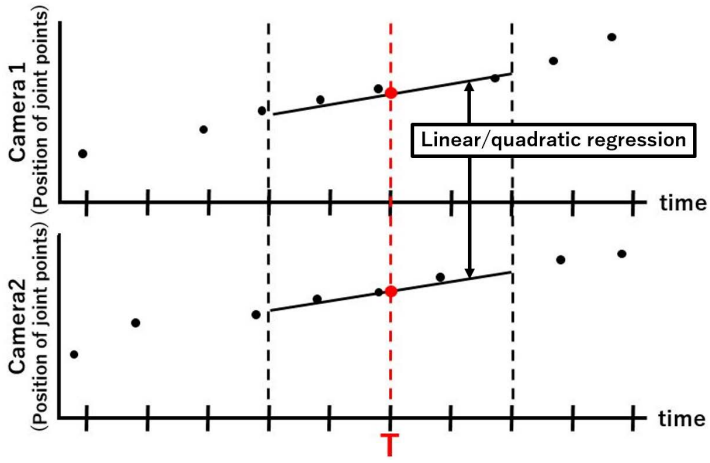

Fig. 4 Temporal Calibration. The value of time T is interpolated from the data group acquired at the surrounding time.

3.2 Camera Calibration

The projection matrix of each camera is required for the 3D pose reconstruction. We thus perform the camera calibration in advance locating many checker patterns in the experimental environment, as shown in Fig. 3 . Positions of the markers in the global coordinate sytem is measured, and they are captured by the cameras to obtain their 2D positions on the image coordinate systems.

3.3 Pseudo-Synchronization

Each camera is connected to a PC (âĂIJNode PCâĂİ) and runs asynchronously. When a moving person is captured from the asynchronous cameras, he/she is captured at slightly different moments among those cameras, so that measured poses are not identical to one another. Also, some data may be missing when saving the image to the PC. To solve these problems, the node PCs are synchronized using an NTP server in advance, and a timestamp is attached to each measurement data. For any moment T, temporally adjacent data points are selected, linear/quadratic regression is applied to those points, and calculate the value at T on the regression curve to obtain the interpolated data, as shown in Fig. 4. T is determined so as to contain three data points within $[-T : T]$; the linear and quadratic regression require at least two and three points, respectively.

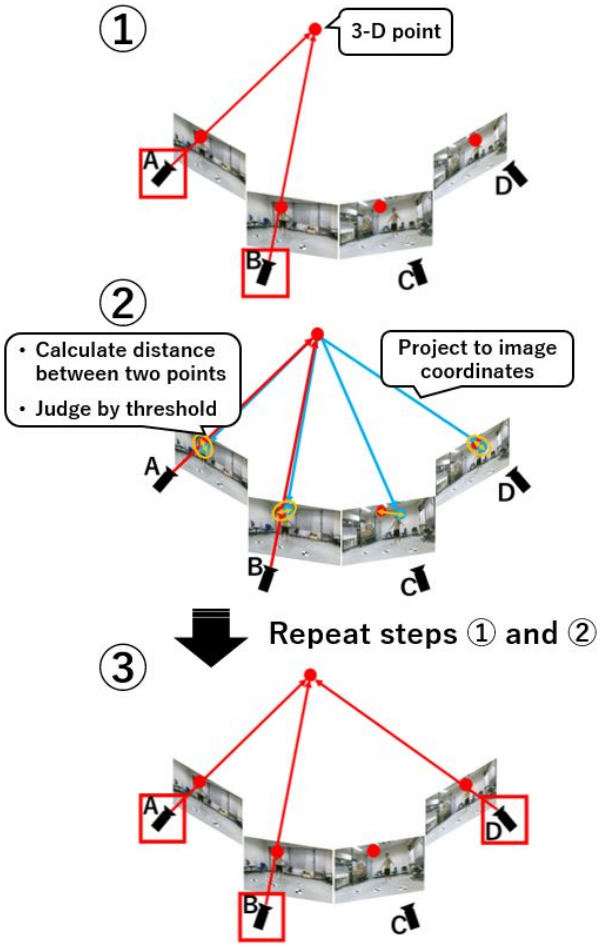

Fig. 5 RANSAC for eliminating inappropriate points.

3.4 3D Pose Estimation (OpenPose Stereo)

Through the processes Sections 3.2 and 3.3, we can obtain the pseudo-synchronous 2D poses from the cameras, which could be used for reconstructing the 3D pose. The 2D poses, However, often lack due to self-occlusion of the person, which prevents us from reconstructing the whole body pose. To overcome this problem, We apply RANSAC to reconstruct the 3D pose.

The specific processing flow is shown below. two cameras are randomly selected as shown in Fig. 5① and 3D coordinates are estimated from the 2D coordinates of the corresponding joint. As shown in Fig. 5②, the 3D point is projected on all image coordinate systems, and the distance between the 2D coordinate estimated by OpenPose and the projected point is calculated. Then, it determines whether it is

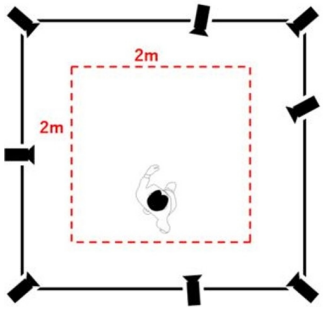

Fig. 6 Camera locations.

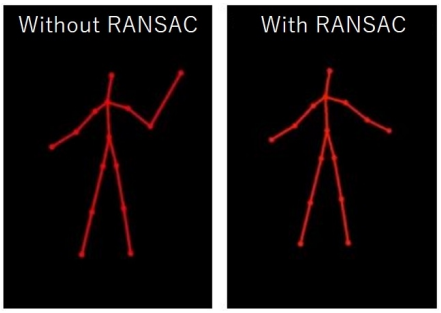

Fig. 7 Effect of RANSAC.

inlier or outlier using the set threshold. At this time, record the number of inliers and the camera selected as the inlier, and repeat the processing of ① and ②. Update the record only if more points in the iteration are chosen as inliers. The update process is repeated several times, and the 3D coordinates are estimated using the 2D coordinates that still remains as the inliers.

4 Experiment

4.1 Experimental Settings

In this experiment, an experimental environment was constructed using eight cameras, which are consumer webcams and so cannot run synchronously. The arrangement of cameras and the person is as shown in Fig. 6. All the cameras captured a participant who walked or ran in the two-meter square area defined around the center of the environment.

In all experiments, the two regression models (linear/quadratic) were applied for generating the pseudo-synchronous data.

4.2 Effect of RANSAC

Fig. 7 shows an example of reconstructed 3D pose. The left and right figures show the results with and without RANSAC, respectively. From these results, we confirmed the effectiveness of the RANSAC process.

4.3 Robustness against Lower Frame Rate

The purpose of this experiment is to evaluate the availability of the proposed method in realistic situations. It is desirable to compare the 3D poses estimated from asynchronous cameras using the proposed method with those from physically synchronous cameras (i.e., ground truth). However, it is difficult because it costs a lot to prepare such special devices, which is actually one of the motivations of this work, as discussed in Section 1. Therefore, instead of using the synchronous cameras for obtaining the ground truth, we captured the scenes by asynchronous cameras but with their highest frame rate (30 fps), and applied the proposed method to estimate 3D poses. We then regarded the 3D poses as the "baseline," and compared the results from the lower frame rate with this baseline. The lower frame rate sequences (15, 10, 6, 5, 3, 2, 1 fps) were generated by simply picking up frames from the highest frame rate sequence.

We also discuss how low frame rate is acceptable for practical use. For this discussion, we adopt a criterion that the average error of 3D pose estimation (i.e., difference from the baseline) should be less than 30 mm.

The following sections describe the results under walking and running scenes.

4.3.1 Walking Scene

Fig. 8 and Fig. 9 show the effect of changing the frame rate. Fig. 8 shows the average errors per frame of the neck and right wrist joints. The error increases as the frame rate decreases. It is also confirmed that while the neck, which fluctuates a little, tends to have small errors, those of the wrist is much larger since it moves faster and larger.

A comparison of the regression models is also important. From Fig. 8 and Fig. 9, it is confirmed the quadratic interpolation has less error than linear interpolation; at 3 fps, while the error was about 100 mm in the linear interpolation, it was just less than 40 mm in the quadratic interpolation.

Fig. 9 shows that the average errors of all the joint points. This result indicates that we should adopt the quadratic interpolation, and that when we would like to estimate 3D poses of a walking person, the frame rate should be no lower than 3 fps.

4.3.2 Running Scene

Fig. 10 and Fig. 11 show the effect of changing the frame rate. It is confirmed that the neck/wrist/average errors were generally larger than those under the walking condition. The quadratic interpolation gave better results even under the running scene as well as the walking scene. Its effect, however, got smaller; the errors by the quadratic interpolation are just similar to (not so better than) those by the linear interpolation, especially in the low frame rate cases (1,2,3 fps). It would be because the running motion is too fast to reproduce from such low frame rate sampling so that whichever interpolation models should never improve the performance.

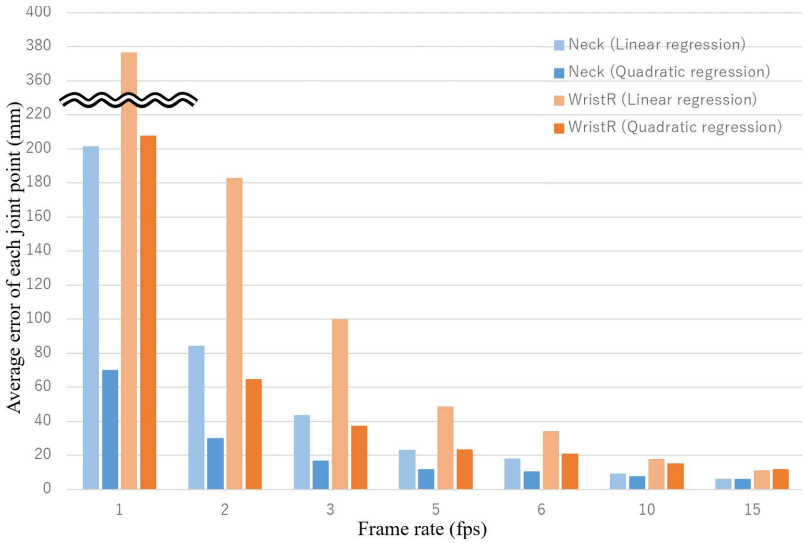

Fig. 8 Average error of each joint point in walking scene.

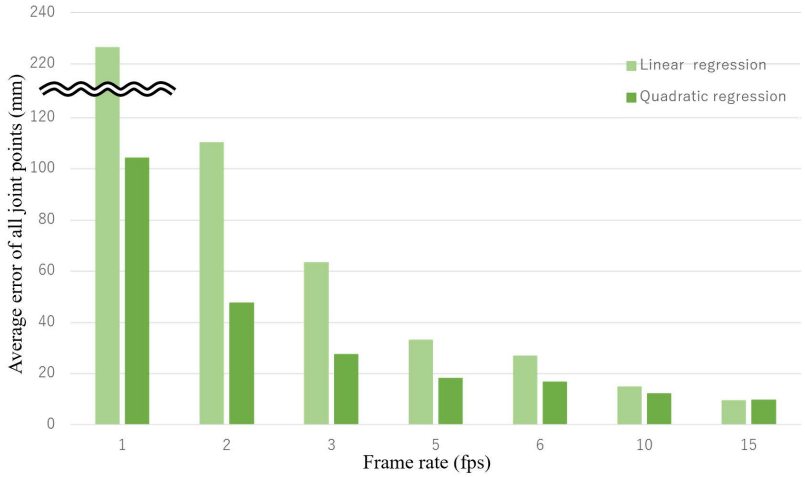

Fig. 9 Average error of all joint points in walking scene.

From the result, we found that the minimum frame rate for the running scene is 10 fps, which is much higher than that for the walking scene.

5 Conclusion

In this paper, we proposed a method for estimating 3D pose using multiple asynchronous cameras (OpenPose Stereo method). In this method, 3D pose estimation

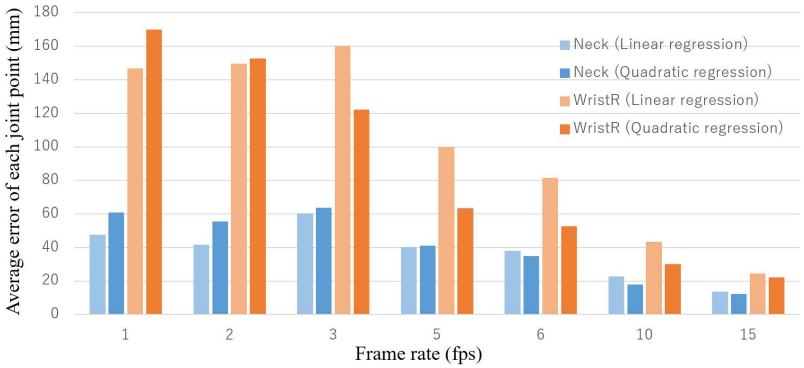

Fig. 10 Average error of each joint point in running scene.

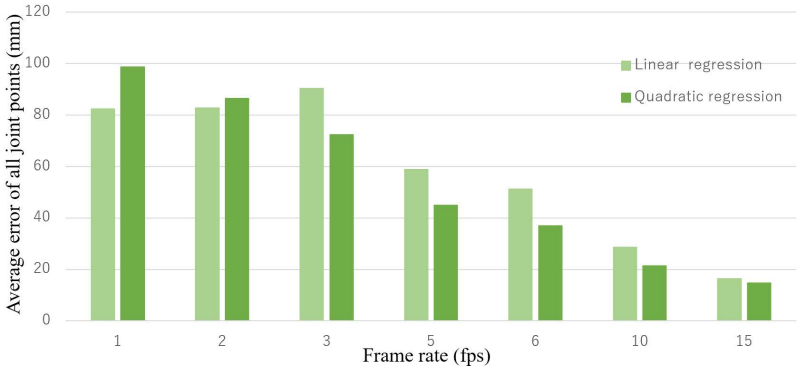

Fig. 11 Average error of all joint points in running scene.

using asynchronous cameras was realized by generating pseudo-synchronous 2D poses from 2D poses that are obtained from actually captured images. First, we captured a moving person with multiple cameras and applied OpenPose to the obtained RGB images to estimate the 2D poses. Then, pseudo-synchronous 2D poses were generated by resampling, and the 3D pose was estimated by triangulation. Triangulation was performed robustly using RANSAC.

In the experiments, we investigated the effect of the proposed method and the effect of the frame rate on the accuracy of pose estimation. It was confirmed that the 3D pose estimation was performed correctly by eliminating inappropriate joint points using RANSAC. The effect of the frame rate on the attitude estimation accuracy was quantitatively evaluated by calculating the error when the frame rate was 30 fps and when the frame rate dropped.

Future work contains improving the method so as that it would not require the camera calibration in advance to realize the motion capture using even unfixed (mobile) cameras.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number JP18H03312.

References

1. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," arXiv:1812.08008, 2018.
2. Z. Cao, T. Simon, S. Wei, Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291-7299, 2017.
3. T. Simon, H. Joo, I. Matthews, Y. Sheikh, "Hand Keypoint Detection in Single Images using Multiview Bootstrapping," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2017.
4. S. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, "Convolutional pose machines," arXiv preprint arXiv:1602.00134, 2016.
5. M. Nakazawa, I. Mitsugami, H. Habe, H. Yamazoe, Y. yagi, "Calibration of multiple Kinects with Little Overlap Regions," Proceedings of the IEEJ Transactions on Electrical and Electronic Engineering, Vol.10, No.S1, pp.S108-S115, 2015.
6. J. Murakami, T. Morimoto, I. Mitsugami, "Gaze and Body Capture System under VR Experiences," Proceedings of the ACM Symposium on Virtual Reality Software and Technology, No.91, 2018.
7. J. Shotten, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images," Proceedings of the IEEE Computer Vision and Pattern Recognition, 2011.
8. R. Girshick, J. Shotton, P. Kohli, A. Criminisi, A. Fitzgibbon, "Efficient Regression of General-Activity Human Poses from Depth Images," Proceedings of the IEEE International Conference on Computer Vision, 2011.
9. J. Martinez, R. Hossain, J. Romero, J. Little, "A simple yet effective baseline for 3d human pose estimation," Proceedings of the IEEE International Conference on Computer Vision, pp. 2640âĂŞ2649, 2017.
10. S. Li, W. Zhang, and AB. Chan, "Maximum-margin structured learning with deep networks for 3d human pose estimation," Proceeding of IEEE International Conference on Computer Vision, pp. 2848-2856, 2015.
11. BX. Nie, P. Wei, SC. Zhu, "Monocular 3d human pose estimation by predicting depth on joints," Proceeding of IEEE International Conference on Computer Vision, 2017.
12. MRI. Hossain, JJ. Little, "Exploiting temporal information for 3d human pose estimation. In The European," Proceedings of the European Conference on Computer Vision, pp. 68-84, 2018.
13. B. Wandt, B. Rosenhahn, "RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7782-7791, 2019.
14. I. Habibie, W. Xu, D. Mehta, G. Pons-Moll, C. Theobalt, "In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10905-10914, 2019.
15. C. Chen, A. Tyagi, A. Agrawal, D. Drover, R. MV, S. Stojanov, JM. Rehg, "Unsupervised 3D Pose Estimation with Geometric Self-Supervision," Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5714-5724, 2019.
16. H. Rhodin, J. SpÃűrri, I. Katircioglu, V. Constantin, F. Meyer, E. MÃijller, M. Salzmann, P. Fua, "Learning Monocular 3D Human Pose Estimation from Multi-view Images," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8437-8446, 2018.
17. X. Chen, K. Lin, W. Liu, C. Qian, X. Wang, L. Lin, "Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10895âĂŞ10904, 2019.
18. M. Kocabas, S. Karagoz, E. Akbas, "Self-supervised learning of 3d human pose using multi-view geometry," arXiv preprint arXiv:1903.02330, 2019.
19. Z. Zhang, "A Flexible New Technique for Camera Calibration," Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1330-1334, 2000.