Video Motion Capture from the Part Confidence Maps of Multi-Camera Images by Spatiotemporal Filtering Using the Human Skeletal Model

Takuya Ohashi, Yosuke Ikegami, Kazuki Yamamoto, Wataru Takano and Yoshihiko Nakamura

Abstract— This paper discusses video motion capture, namely, 3D reconstruction of human motion from multi-camera images. After the Part Confidence Maps are computed from each camera image, the proposed spatiotemporal filter is applied to deliver the human motion data with accuracy and smoothness for human motion analysis. The spatiotemporal filter uses the human skeleton and mixes temporal smoothing in two-time inverse kinematics computations. The experimental results show that the mean per joint position error was 26.1mm for regular motions and 38.8mm for inverted motions.

I. INTRODUCTION

Motion capture technology is widely used for many applications, in particular with human motion, such as animation, video game production, sports training, biomechanical analysis, rehabilitation, medical diagnosis, and research of human motion modeling and recognition [1][2][3].

The technology has three typical methods for measurements, namely, passive optical, active optical, and gyro-accelerometer measurements. The spatiotemporal accuracy is approximately 1mm and 1kHz for the typical human scale environments in the case of the high-end passive optical system [4][5], and getting even higher by the advancement of imaging device technology. The passive optical method imposes the least constraint to a human subject, but still needs approximately 40 spherical reflective markers attached on the body. Other methods also have similar or even stronger constraints.

To solve this problem, image-based motion capture which imposes no constraint to the human subjects is demanded. The studies on RGB-D vision show promising results for human motion recognition using Kinect [6][7] and the other systems with laser projection or scanning depth sensors [8][9]. Recent studies on single RGB vision [10][11][12][13] show remarkable results where, taking an advantage of the large RGB image data set and the power of deep neural networks, the learning based methods for human motion recognition have been developed.

Although image-based human motion recognition is considered a 2D problem and rather tolerant in spatial accuracy and temporal discontinuity, motion capture is a 3D reconstruction problem and requires spatiotemporal accuracy and smoothness for velocity and acceleration analysis.

In this paper, we report about our study on video motion capture, which means image-based 3D human motion reconstruction with spatiotemporal accuracy and smoothness closer to passive optical motion capture. We use synchronized multiple cameras to record video images of a human subject from different directions. The likelihood of joint positions are computed as the Part Confidence Maps (PCM) [14][15] in each video image and provided by the colors of heatmap with the labels of individual joints by OpenPose [16]. Our initial study showed that the simple least square solution of the 3D position of each joint from the multi-camera images neither resulted in smooth 3D joint movements nor maintained link length between adjacent joints.

We propose to use the human skeletal model for reconstructing 3D human motion by spatiotemporal filtering of joint movements. We also discuss to enhance function of the PCM for rather difficult human postures. The proposed video motion capture was implemented and tested through various experiments. The results of evaluation are shown based on the reconstructions made by two methods, namely, the video motion capture and the passive optical motion capture.

II. VIDEO MOTION CAPTURE FROM MULTI-CAMERA IMAGES

A. *Flow of the Video Motion Capture*

Proposed video motion capture can be described as the iterative update of human skeletal model's joint angles based on multi-camera images which surround a human subject like Fig.1. Note that in this paper the skeletal model means a virtual open tree-structure kinematic chain with 40 degrees of freedom (DOF) as shown in Fig.2 [a]. To fulfil this update, the weighted markers which are virtually generated in the 3D space are computed as shown in Fig.2 [b] based on the subject's probable joint positions. By minimizing the errors of virtual markers' positions and skeletal model's corresponding joint positions, the skeletal model's joint positions at the time are obtained. To solve this minimisation problem, inverse kinematics (IK) computation is adapted, and by repeating above process, the video motion capture is realized.

The flow chart of the method is shown in Fig.3. Note that the camera calibration must be done before the computation.

The camera parameters such as the focal length, the center of image, and the lens distortion enable restoration of distorted images and provide the transformation matrices $\mathcal{P}_i$ that maps an arbitrary 3D point to the corresponding pixel location on the ith camera image plane [17][18]. Other main blocks in the chart are explained in the following subsections.
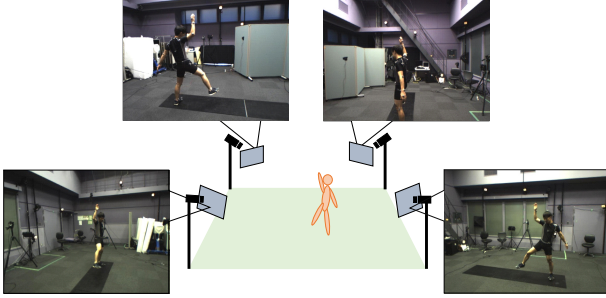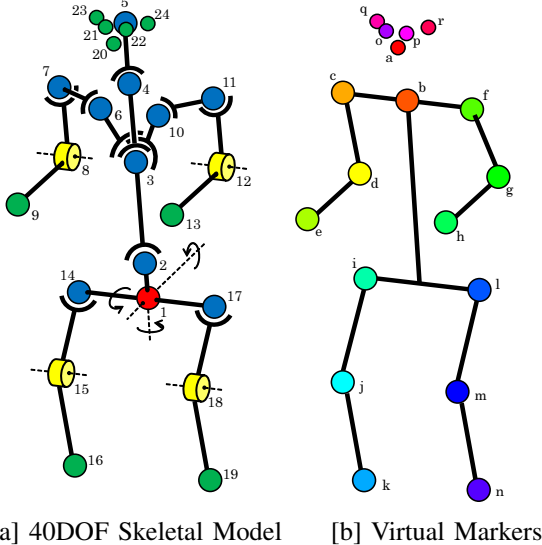

Fig. 1: Measurement Environment

[a] 40DOF Skeletal Model   [b] Virtual Markers

Fig. 2: Correspondence of Human Skeletal Model and Virtual Markers. In [a], joints color indicates as 6DOF (red), 3DOF (blue) and 1DOF (yellow)

B. Computation of the Part Confidence Maps

The OpenPose [16] provides the Part Confidence Maps (PCM) [14] and the Part Affinity Fields (PAF) [15] of multi-person's keypoints on a 2D image in real time using the trained Convolutional Neural Networks (CNN). The keypoints are 18 consisting of joint positions like shoulders, elbows, wrists, hips, knees, ankles and a neck, and the other feature points like eyes, ears, and a nose. Though the recent version of OpenPose includes hands, feet, and face, we used only the above 18 keypoints.

The PCM shows by its temperature the distribution of likelihood of the keypoint, namely, in each pixel location, the probability of the keypoint is expressed as the continuous
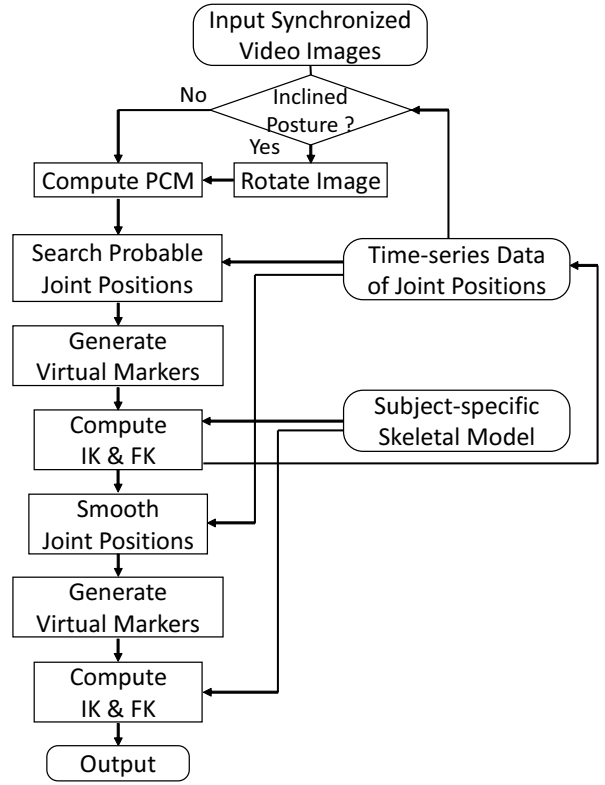

Fig. 3: Flow Chart of the Video Motion Capture

value from 0 to 1. The centroid of the PCM may be considered as the position of the corresponding keypoint on the 2D image. Note that we use the PCM not its centroids for the 3D human motion reconstruction. The PAF shows the connectivity of the keypoints of a single person. In this paper, we do not use the PAF though it is a powerful means to identify individuals of multi-person on the image. An example of input image [a], PCM [b], PAF [c], and the centroids of PCM superimposed on the image [d] are shown in the Fig.4
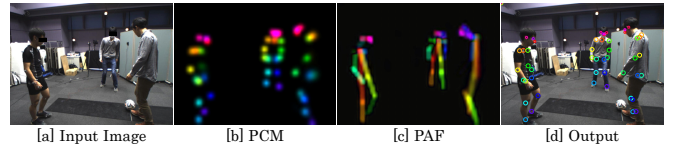

[a] Input Image   [b] PCM   [c] PAF   [d] Output

Fig. 4: Multi-Person Keypoints Detection Using OpenPose

C. Initialization of Skeletal Model and Joint Positions

In order to compute IK, the skeletal model's adjacent joints must be connected by a constant-length link and the link length must be calculated according to the human subject. Furthermore, since proposed motion capture is based on iterative computation, it is reasonable to calculate the skeletal model's initial joint positions before the computation.

The link lengths of the skeletal model are identified for the subject by using the centroids of PCM on the images

from multiple cameras as schematically shown in Fig.5. Initialization is confirmed when the centroids of every keypoint from the cameras show a 3D agreement. And simultaneously, skeletal model's initial joint positions are calculated.
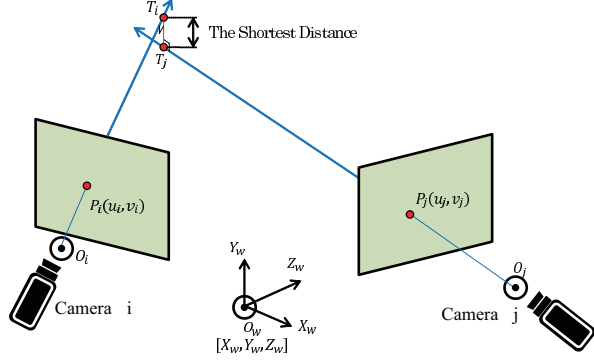

Fig. 5: 3D Position Calculation from Pixel Location

D. Searching Probable Joint Positions

Let $^tJ^n$ and $^tQ^n$ represent the 3D joint position and the joint angle of nth joint at t time-frame. We discretize the 3D space with a unit distance s, and consider the lattice of $(2k+1) \times (2k+1) \times (2k+1)$ with $^tJ^n$ at the center. We then search for $^{t+1}J^n$ on the lattice. The search space $^t\mathbb{J}^n$ is represented by

$$^t\mathbb{J}^n := \left\{ ^tJ^n + s \begin{bmatrix} a \\ b \\ c \end{bmatrix} \;\middle|\; -k \le a,b,c \le k \right\} \quad (1)$$

$$k : \text{constant positive integers}$$

$$a,b,c : \text{integers}$$

A point on the lattice in $^t\mathbb{J}^n$ is described as $^tJ^n_{a,b,c}$. It is considered that the point with the highest score from summation of multi-camera PCM values is the most probable point of the joint existence. Let n_c be the number of cameras and the function $^{t+1}\mathcal{S}^n_i$ be the PCM value that of a corresponding pixel positionof nth joint on ith camera image at $t+1$. Then, the predicted joint position at $t+1$ can be calculated as follows:

$$^{t+1}J^n_{pred} = \operatorname*{arg\,max}_{-k \le a,b,c \le k} \sum_{i=1}^{n_c} {}^{t+1}\mathcal{S}^n_i(\mathcal{P}_i(^tJ^n_{a,b,c})) \quad (2)$$

At this predicted joint position, virtual marker is generated.

E. Computing Inverse Kinematics

Since the joint position $^{t+1}J^n$ is a function of joint angle ^{t+1}Q expressed in forward kinematics (FK), ^{t+1}Q can be obtained by solving following optimization problem [19]:

$$^{t+1}Q = \operatorname*{arg\,min} \sum_{n}^{n_j} \frac{1}{2} {}^{t+1}W^n ||^{t+1}J^n_{pred} - {}^{t+1}J^n||^2 \quad (3)$$

Note that n_j is the number of joints and the constant weight $^{t+1}W^n$ is obtained from the PCM score at the predicted joint position as follows.

$$^{t+1}W^n = \sum_{i=1}^{n_c} {}^{t+1}\mathcal{S}^n_i(\mathcal{P}_i(^{t+1}J^n_{pred})) \quad (4)$$

F. Smoothing Joint Positions with Maintaining Link Length

Human movement is smooth and differentiable by time, but acquisition of PCM and IK computation does not use time-series data, so the output joint data is not to be smoothed. To solve this problem, joint position is smoothed by introducing IIR low-pass filter. However, the time-invariant parameters of skeletal model, link length, is changed every time-frame by the smoothing.

In proposed method, the position of the joint after the adaptation of the low-pass filter is set as the virtual marker of each joint, and the optimization by IK computation is performed again. This makes it possible to make use of the smoothness of the temporal change of the joint position under the skeletal condition that each link length is invariant.

III. Experiments

A. Comparison with Optical Motion Capture

Human subject's motion was synchronously recorded by 4 RGB cameras (FLIR Corporation, Grasshopper3) at 60Hz with a 1024×768 pixel resolution and measured by the optical motion capture system with 17 infra-red cameras (Motion Analysis Corporation, Eagle and Raptor-4) at 200Hz with 41 markers. Fig.6 shows the positions of the attached markers. Ground truth of joint positions was calculated by these measured markers.

In this experiment, 9 motions were measured; Walk, Dance, Spin Jump, Football Juggling, Golf, Matrix, Handstand, Cartwheel, and Break Dance.
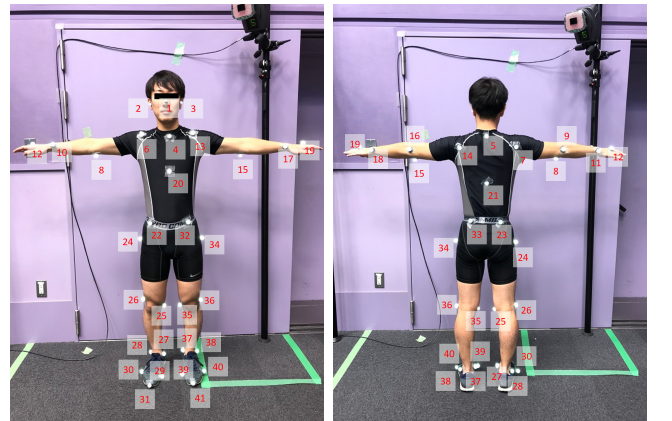

Fig. 6: Positions of the Attached Markers

Fig.7 shows representative frame of each motion. Top images represent input images. Estimated joint positions from the video motion capture are drawn as circles. Bottom images represent 3D pose. Joint positions from the video

motion capture are drawn as stick model and joint positions from optical motion capture are drawn as bone model.
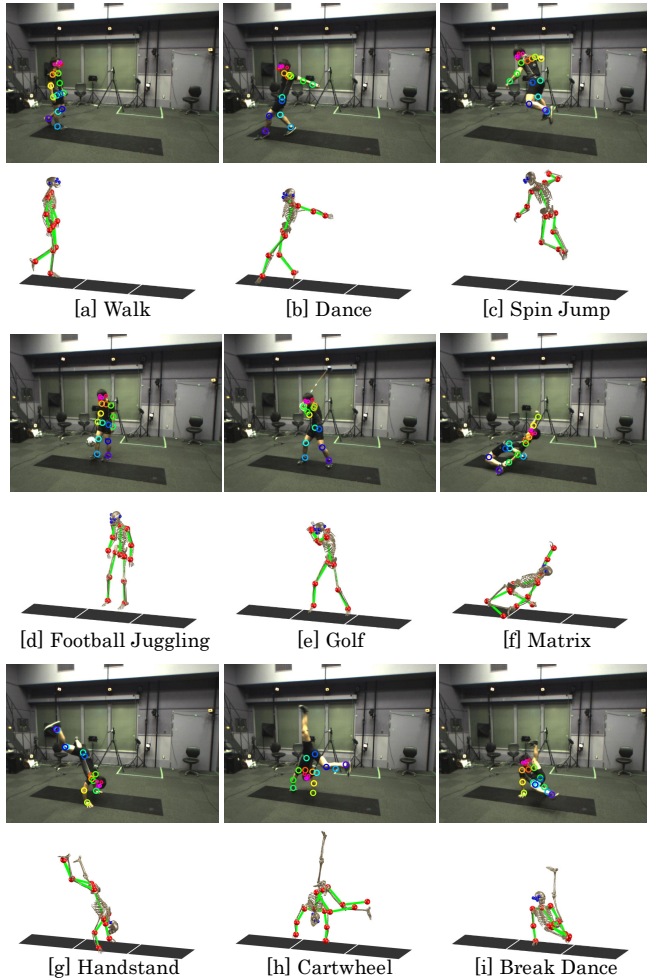

Fig. 7: Results of the Video Motion Capture

[a] Walk   [b] Dance   [c] Spin Jump

[d] Football Juggling   [e] Golf   [f] Matrix

[g] Handstand   [h] Cartwheel   [i] Break Dance

Table.I shows the results of the mean per joint position error (MPJPE) of each motion and percentage of correct keypoints in 3D (3D-PCK) when all the joints are in the images. UpperBody consists of neck, shoulders, elbows, and wrists. LowerBody consists of hips, knees, and ankles. Head consists of nose and ears.

From the results, in the regular motions such as Dance, Spin Jump, and Matrix, proposed method can estimate joint positions properly and achieves the accuracy of about 30mm in MPJPE. Fig.8 shows MPJPE & PCM score change in Football Juggling motion which is the smallest in MPJPE. Note that PCM score is the value of $^{t+1}W^n$ defined in Eq.4. In every frame, proposed method achieves 25mm in MPJPE and PCM score keeps high.
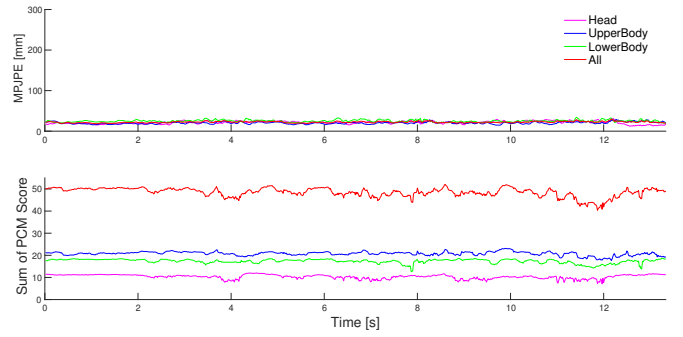

Fig. 8: MPJPE & PCM Score, Football Juggling

However, in inverted motions such as Handstand and Cartwheel, proposed method estimates the joint positions with larger errors, especially in Lowerbody. Fig.9 shows MPJPE & PCM score change in Handstand motion which is the largest in MPJPE. As the body tilts, MPJPE became larger and PCM score getting worse.
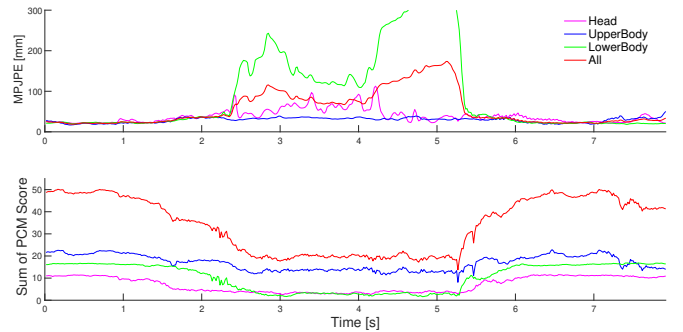
Fig. 9: MPJPE & PCM Score, Handstand

This problem is caused by the accuracy of PCM. Since OpenPose was created by supervised learning, PCM computation is greatly affected by the training data, and it can be considered that training data did not contain a sufficient amount of inverted posture. Fig.10 shows the PCM change by the input image rotation. By inputting rotated image according to the tilt of the human subject, better PCM can be computed.

[a] Input Image   [b] PCM of Right Ankle

Fig. 10: PCM Change by Input Image Rotation

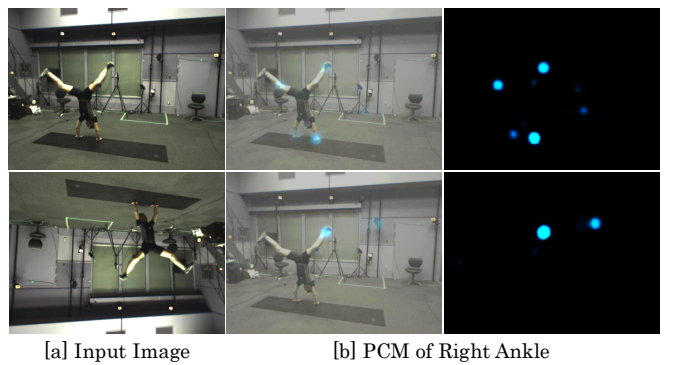TABLE I: Accuracy of Video Motion Capture

		Walk	Dance	Spin Jump	Football Juggling	Golf	Matrix	Handstand	Cartwheel	Break Dance	Total
MPJPE [mm]	Head	19.9	25.4	29.8	21.3	22.9	32.4	38.5	37.2	33.1	26.9
	UpperBody	29.7	29.3	31.4	20.3	32.4	28.5	29.1	34.3	32.4	29.6
	LowerBody	26.0	23.0	23.7	25.0	18.4	30.7	96.0	81.7	57.4	33.3
	Total	26.5	26.2	28.2	22.3	25.3	30.1	56.0	52.6	41.9	30.5
3D-PCK [%] @ 50mm	Head	100	99.8	95.4	100	99.8	88.7	79.7	85.6	88.6	95.9
	UpperBody	99.3	93.0	91.8	99.9	87.0	92.5	89.4	78.6	90.1	92.1
	LowerBody	98.8	97.3	94.5	99.0	100	92.2	72.1	87.0	82.4	94.2
	Total	99.3	95.9	93.5	99.6	94.3	91.7	81.1	83.1	86.9	93.6
3D-PCK [%] @ 100mm	Head	100	100	100	100	100	100	98.5	93.7	99.4	99.6
	UpperBody	100	99.8	99.9	100	98.6	100	100	98.0	98.8	99.5
	LowerBody	100	100	99.5	100	100	100	81.4	90.1	92.4	97.8
	Total	100	99.9	99.8	100	99.4	100	92.7	94.2	96.5	98.9
3D-PCK [%] @ 150mm	Head	100	100	100	100	100	100	99.9	96.3	100	99.8
	UpperBody	100	99.9	99.9	100	99.4	100	100	100	99.3	99.8
	LowerBody	100	100	99.8	100	100	100	86.2	92.0	95.3	98.4
	Total	100	99.9	99.9	100	99.8	100	94.8	96.3	97.9	99.3

B. Effect of Input Image Rotation

[a] Handstand [b] Cartwheel [c] Break Dance
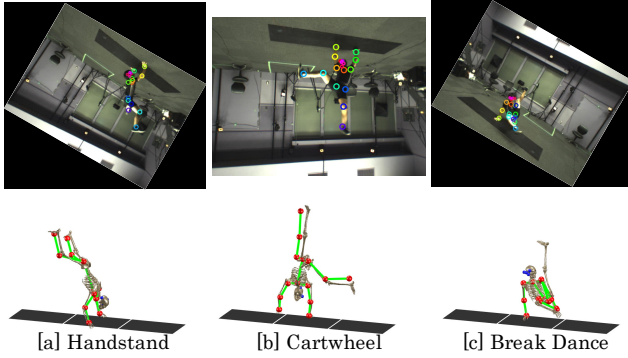
Fig. 11: Results of Video Mocap with Image Rotation

TABLE II: Accuracy Change by Image Rotation

		Handstand	Cartwheel	Break Dance
MPJPE [mm]	LBody	43.4 (-52.6)	48.5 (-33.2)	51.4 (-6.0)
	Total	36.3 (-19.6)	40.2 (-12.5)	39.6 (-2.3)
3D-PCK[%] @ 50mm	LBody	76.0 (+3.9)	85.6 (-1.4)	82.3 (-0.1)
	Total	82.5 (+1.4)	82.4 (-0.7)	86.9 (0)
3D-PCK[%] @ 100mm	LBody	92.2 (+10.8)	93.0 (+2.9)	93.0 (+0.6)
	Total	96.8 (+4.4)	95.3 (+1.1)	96.7 (+0.2)
3D-PCK[%] @ 150mm	LBody	95.9 (+9.7)	96.1 (+4.1)	95.9 (+0.6)
	Total	98.4 (+3.6)	98.0 (+1.7)	98.2 (+0.3)

To improve the MPJPE in inverted motion, we propose computing LowerBody PCM according to the tilt of trunk. The algorithm is as follows.

From the skeletal model's neck and hip joint position at t frame, the tilt of trunk at t is calculated by each camera image. According to the tilt, input image at $t+1$ is rotated and then LowerBody's PCM are computed. PCM of the others are computed from normal input image. Using these PCM, virtual markers in 3D space are generated and IK computation is performed illustrated as above.

The results of incorporating the input image rotation method into Handstand, Cartwheel, Break Dance are shown as follows. Fig.11 shows that joint position estimation became possible even in inverted motion such as Handstand & Cartwheel due to the input image rotation. However, in Break Dance, no significant change appeared. This can be attributed to the difference between whether the axis of rotation of each motion is perpendicular or parallel to the ground.
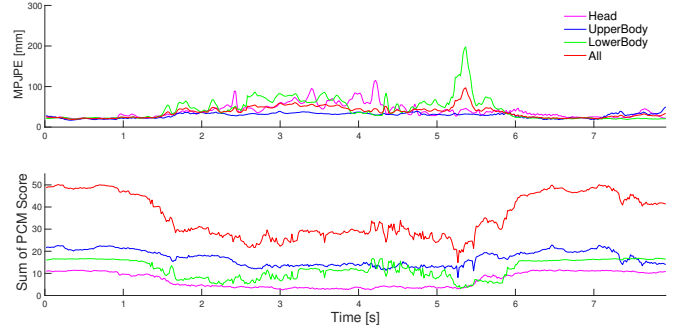

Fig. 12: MPJPE & PCM Score: Handstand with Rotation

To summarize the above results, as shown in the Table.III, we achieved accuracy of 26.1mm of regular motion and 38.8mm of inverted motion in MPJPE. Also, we achieved accuracy of 99.8 % for regular motions, and 96.4 % for inverted motions in 3D-PCK at 100mm. Proposed method is usable regardless of human subject's cloth and environment as Fig.13.

TABLE III: Summary of Accuracy of Video Motion Capture

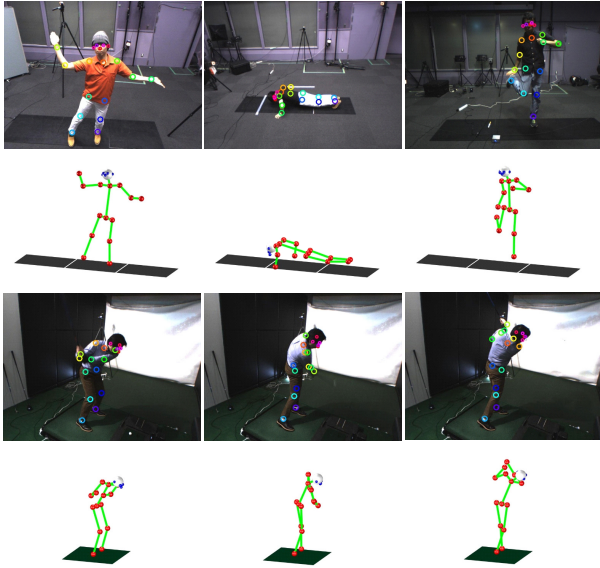|		Regular Motion	Inverted Motion with Rotate
MPJPE [mm]	Head	24.8	35.9
	UpperBody	29.0	31.8
	LowerBody	23.4	48.3
	Total	26.1	38.8
3D-PCK [%] @ 50mm	Head	98.4	84.9
	UpperBody	93.3	87.1
	LowerBody	97.5	81.2
	Total	95.8	84.5
3D-PCK [%] @ 100mm	Head	100	97.8
	UpperBody	99.7	98.9
	LowerBody	99.9	92.7
	Total	99.8	96.4
3D-PCK [%] @ 150mm	Head	100	99.3
	UpperBody	99.9	99.7
	LowerBody	99.9	96.0
	Total	99.9	98.2

Fig. 13: Results on Real Environments

IV. CONCLUSION

The conclusion of this paper is as follows:

1) An approach of video motion capture was proposed towards image-based 3D human motion reconstruction from multi-camera images using the Part Confidence Maps.

2) A method of spatiotemporal filtering was developed using the human skeletal model for the 3D reconstruction of human motion by mixing temporal smoothing in between two-time inverse kinematic resolutions.

3) The accuracy of the Part Confidence Map was improved using the rotated images for inverted motions such as a handstand and a cartwheel.

4) The experimental results of video motion capture showed that the mean per joint position error was 26.1mm for regular motions and 38.8mm for inverted motions. When the threshold was set at 100mm as the maximum error of all joint positions for the whole

motion, the success rate was 99.8 % for regular motions and 96.4 % for inverted motions.

V. ACKNOWLEDGEMENTS

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

[8] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart. A Layered Approach to People Detection in 3D Range Data. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[9] A. Dewan, T. Caselitz, G. D. Tipaldi, and W. Burgard. Motion-based detection and tracking in 3D LiDAR scans. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[10] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D Human Pose Estimation Using Transfer Learning and Improved CNN Supervision. *The Computing Research Repository*, abs/1611.09813, 2016.

[11] 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.

[12] 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.

[13] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional Human Pose Regression. *The Computing Research Repository*, abs/1704.00159, 2017.

[14] 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.

[15] 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.

[16] OpenPose. https://github.com/CMU-Perceptual-Computing-Lab/openpose.

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

[18] Itseez. Open Source Computer Vision Library. https://github.com/itseez/opencv.

[19] 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.