3D Human Pose Estimation of a Partial Body from a Single Image and Its Application in the Detection of Deterioration in Sitting Postures

Oky Dicky Ardiansyah Prima and Kazuki Hosogoe
Graduate School of Software and Information Science, Iwate Prefectural University
152-52 Sugo Takizawa, Japan
email: prima@iwate-pu.ac.jp, g231r026@s.iwate-pu.ac.jp

***Abstract*—Three-dimensional (3D) human pose estimation has been used in a wide range of fields, including motion analysis in sports and rehabilitation, modeling in Computer Graphics (CG) production for movies and games, and input interfaces. Recently, 3D human pose can be estimated with high accuracy only from a single image using a neural network model. However, depending on the camera's position and shooting angle, some joints may be occluded, thus reducing the accuracy of the overall joint estimation. In this study, we experimentally constructed a neural network model for 3D human pose estimation based on a single image and evaluated the difference in accuracy of the pose estimated by the model constructed for the partial joints of the body and the whole-body joints. The dataset used for the experiment was Human 3.6M and a human pose dataset created from an RGB-D camera for this study. The results confirmed that the model built based on the upper-body joints of the body had higher accuracy than that for the whole-body joints at estimating the posture of the upper body. Finally, we demonstrated that 3D human pose can be used to detect the deterioration in sitting postures, which can suggest that the technology is effective in improving various postures in daily life.**

Keywords-pose-estimation;3D human pose; body joint; computer vision.

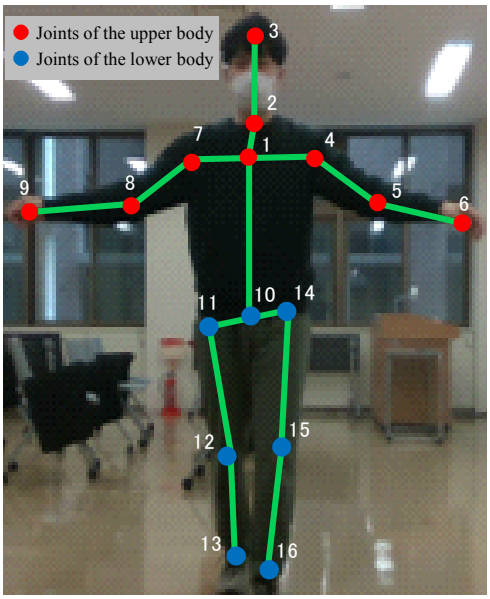


Figure 1. Sixteen joints extracted from the dataset for this study.

I. INTRODUCTION

Three-Dimensional (3D) human pose estimation is the problem of locating the position of human joints in an image or video. Estimating human pose is an important problem that has received a lot of attention in the field of computer vision over the past few decades. Recently, improvements in 3D human pose estimation based on neural networks have led to higher accuracy and higher frame rates, as well as lower prices and simpler use of measurement equipment for human pose estimation [1][2]. As a result, 3D human pose estimation is being used in a wider range of fields, and the scope of its application is expanding not only to large-scale applications by companies and research institutions but also to personal use and other small-scale applications. A typical example of the use of human pose estimation technology at the individual level is video production using virtual characters whose movements are synchronized in real time. The use of such virtual characters with synchronized movements has been rapidly expanding in the entertainment field.

3D human pose information can be used for motion analysis to evaluate the effectiveness of rehabilitation. Using this information, it becomes convenient to measure the range of motion of the patient's joints and to perform self-rehabilitation at home. Experiments conducted by Prima et al. (2019) indicated that 3D human pose estimation from a single image is more advantageous for estimating semi-occluded joint locations than those estimated by depth cameras [3].

Existing studies on 3D human pose estimation have mainly focused on developing and verifying methods assuming that the whole-body joints can be captured. However, depending on the capturing environment and angle, a part of the body may be occluded and thus only some parts of the body can be captured. For example, when a person is sitting at a desk that is fully covered and photographed from the front, the lower half of the body is hidden by the desk and only the upper half of the body is captured. In such a situation, if 3D human pose estimation is performed for all joints of the body, the missing information of some joints may affect the detection of other joints and reduce the estimation accuracy.

This study aims to compare the accuracy of human 3D human pose estimation based on a single image using a model consisting of the whole-body joints and a model consisting of partial joints without occlusion, and to verify the effectiveness of the model corresponding to partial joints when a part of the body is occluded. As an application, a technique for detecting deterioration in sitting postures using 3D human pose estimation is introduced.

 ISBN: 978-1-61208-872-3

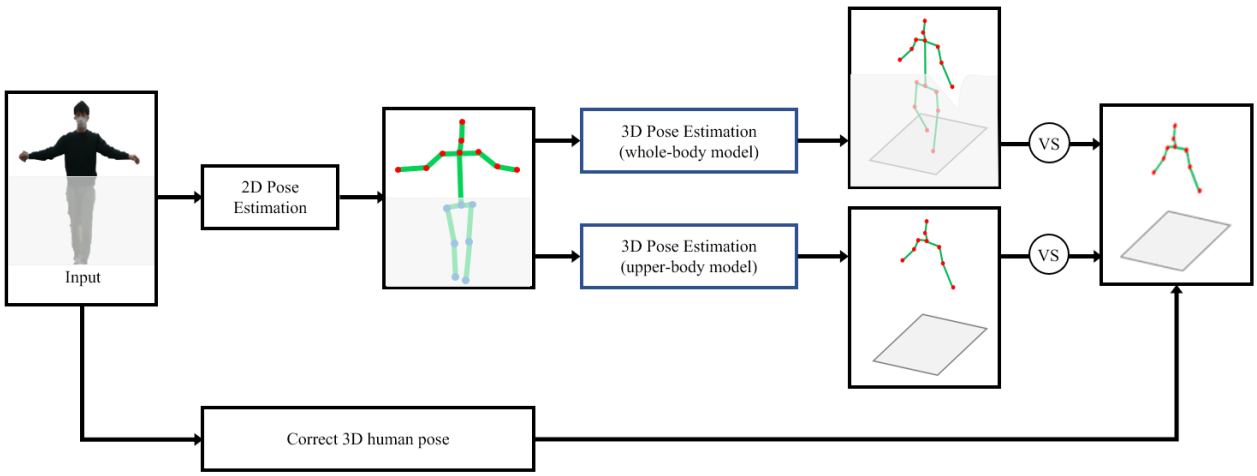


Figure 2. The evaluation flow of the 3D human pose acquired with the whole-body model and the upper-body model, respectively.

This paper is organized as follows. In Section II, we describe the 3D human pose models built for this study. Section III describes the method used to evaluate the accuracy of the partial 3D human pose. In Section IV, we present our results evaluated using the public dataset and the dataset created by an RGB-D camera. Section V introduces a technique for detecting deterioration in sitting postures using 3D human pose estimation. Finally, Section VI summarizes the results of this study and discusses future perspectives.

II. Building 3D Human Pose Models

In this study, we used the Human3.6M [4] to build 3D human pose models. Human3.6M is a 3D human pose dataset consisting of 3.6 million human poses and corresponding images, recorded from the performances of five female and six male subjects. This dataset consists of 2D joint positions, 3D joint positions, RGB images, time-of-flight (depth) data, and 3D body scan data from four different viewpoints measured by a high-speed motion capture system. Sixteen joints (Figure 1), were finally extracted from the dataset, and used in the evaluation of 3D human pose in this study. There are nine joints in the upper body and seven in the lower body.

We chose the 3D baseline method [1] to build our 3D human pose models because it can achieve low error rates on 3D human pose estimation using a relatively simple deep feed-forward neural network. The network expands the dimension of the input data to 1024 using the initial weights by He et al.'s (2015) [5], and then performs a series of batch normalization, Rectified Linear Unit (RELU), dropout rate of 0.5, and a residual connection. Following [1], subjects 1, 5, 6, 7 and 8 were used for training, and subjects 9 and 11 for evaluation. Two models, i.e., a 3D human pose model trained using joints of the whole body, and a model trained using only joints of the upper body were built for this study. Hereinafter, the former is referred to as the "whole-body model" and the latter as the "upper-body model."

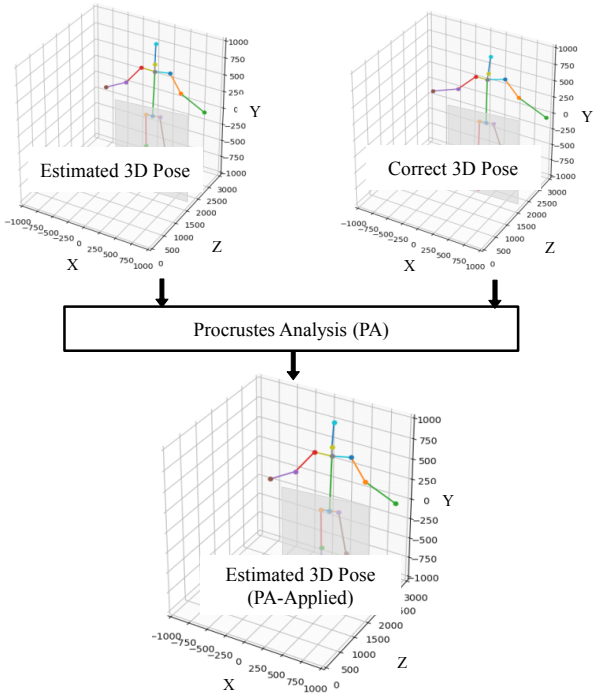


Figure 3. Procrustes analysis applied to the estimated 3D pose.

III. Evaluation of Partial 3D Human Poses

Our evaluation process goes as follows (Figure 2). After extracting the two-dimensional (2D) human pose from the image, we obtain the 3D human pose estimated by each model. For the full-body model, the 3D joints of the whole body are acquired, whereas for the upper-body model, the 3D joints of the upper body are acquired.

We then compare the obtained 3D human poses with the correct 3D human poses of the upper body and calculate the

 ISBN: 978-1-61208-872-3

TABLE I. ERRORS IN UPPER BODY POSES IN THE HUMAN3.6M ESTIMATED BY THE WHOLE-BODY AND THE UPPER-BODY MODELS.

No.	Joints	Model [mm]	
		Whole-Body	Upper-Body
1	Neck	162.79	51.24
2	Nose	38.39	26.71
3	Head	80.75	45.65
4	Left Shoulder	81.05	37.39
5	Left Wrist	91.10	45.13
6	Left Elbow	109.46	69.94
7	Right Shoulder	72.66	35.43
8	Right Wrist	89.63	46.64
9	Right Elbow	105.65	69.50
Mean (M)		92.387	47.514
Standard deviation (SD)		33.5448	14.5406

TABLE II. ERRORS BETWEEN THE 3D POSES OF THE UPPER BODY ESTIMATED BY THE WHOLE-BODY AND THE UPPER-BODY MODELS AND BY THE RGB-D CAMERA.

No.	Joints	Model [mm]	
		Whole-Body	Upper-Body
1	Neck	114.70	74.87
2	Nose	181.63	145.16
3	Head	96.70	**122.22**
4	Left Shoulder	103.78	91.44
5	Left Wrist	102.63	92.35
6	Left Elbow	148.85	107.67
7	Right Shoulder	107.51	**113.34**
8	Right Wrist	102.86	**108.82**
9	Right Elbow	158.45	111.21
Mean (M)		124.123	107.453
Standard deviation (SD)		30.7033	20.1145

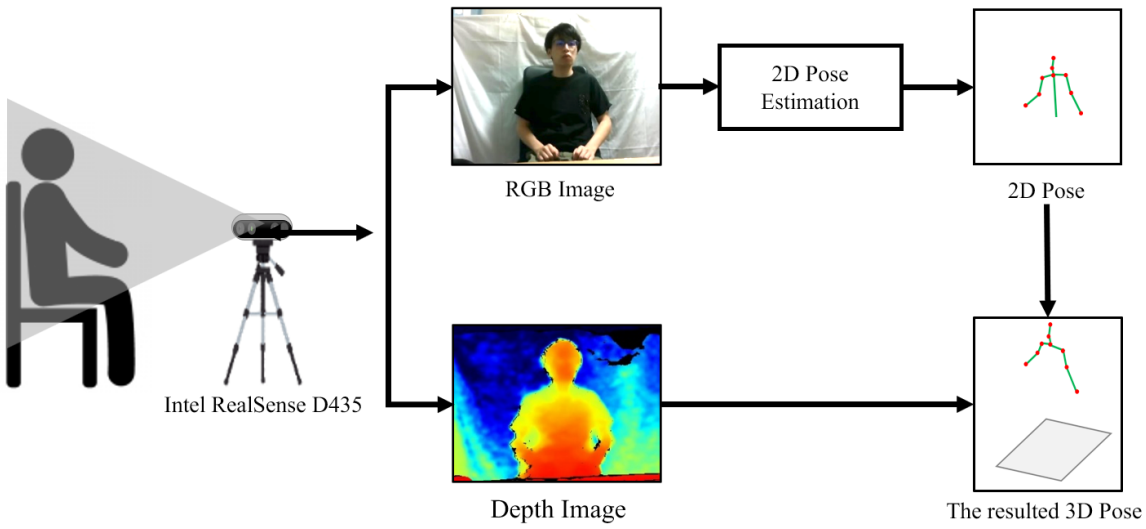


Figure 4. 3D human pose data generation using an RGB-D camera (Intel RealSense D435).

accuracy of the pose estimation by each model. Finally, we evaluate the difference in the accuracy of 3D human pose estimation between the model for the whole body and the model for the upper body in the situation where only the upper body is captured.

To find differences in poses, Procrustes Analysis (PA), a shape-preserving Euclidean transform, is applied to the estimated 3D human pose with reference to the correct one (Figure 3). This analysis eliminates the variation in movement, rotation, and scaling between the pose data while preserving the shape.

IV. RESULTS

To evaluate the difference in accuracy between the whole-body and the upper-body models, we randomly selected 548,800 human poses from the Human3.6M, which were not used for training and validation of these models. Table I shows the estimation errors of the upper body poses by each model. The accuracy of the estimation by the upper-body model is high for all joints. Overall, the accuracy of pose estimation significantly improved by about 45mm when the upper-body model was used to estimate the pose of the upper body (M = 92.387, SD = 33.5448, $t(8) = 4.99$, $p < .001$).

In order to confirm the results of estimating the 3D human pose in a real world, we further evaluated these models with

 ISBN: 978-1-61208-872-3

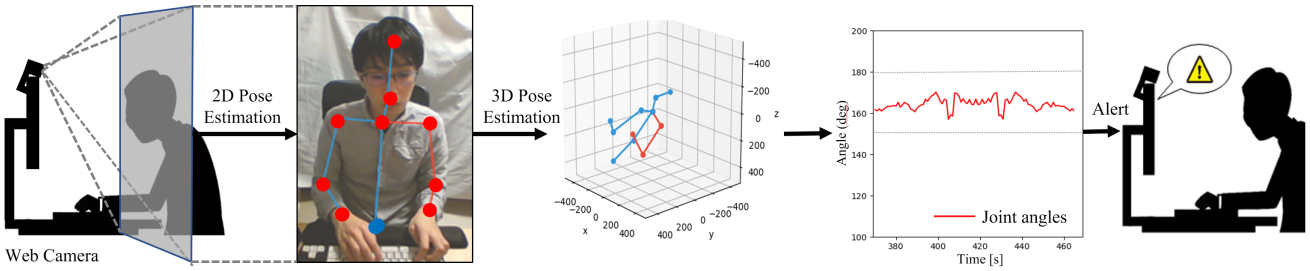
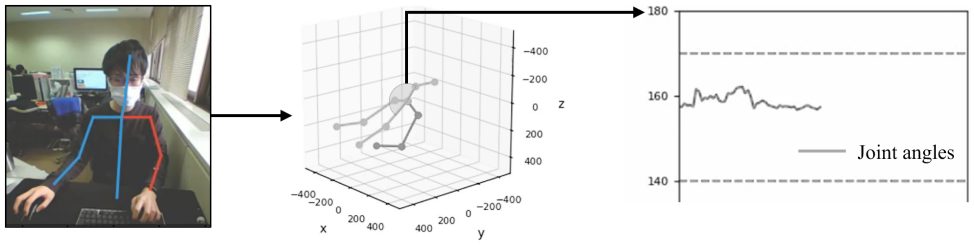
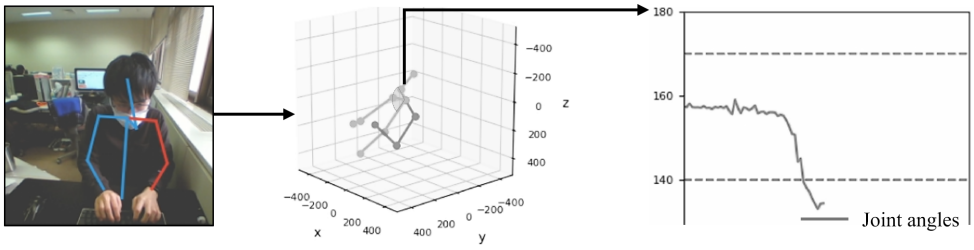


Figure 5. The detection of deterioration in sitting postures based on 3D human pose estimation.



(a) Good sitting postures.



(b) Deteriorated sitting postures.

Figure 6. The difference in the angle formed by the nose, neck, and pelvis in good and deteriorated sitting postures.

the results of measurements using an RGB-D camera. For this study, we generated 7,350 3D human pose data from a subject using the Intel RealSense Depth Camera D435. The subject was seated about 1m from the camera and moved his hands and body in various ways as long as his body did not move out of frame from the camera.

To generate 3D human pose data from the D435, we first apply geometric corrections to the RGB and depth images from the camera to match the coverage of both images (Figure 4). For RGB images, the Stacked Hourglass Network [6] is used to estimate the 2D human pose. By superimposing the 2D human pose information with the depth image, the depth data of each joint is extracted to generate 3D human pose data. However, since the depth data is affected by noise which makes it difficult to obtain the accurate depth of the joint, we filtered this data using a five-frame moving average in time series. The camera's resolution was set to 640x480 pixels and the frame rate to 30fps.

Table II shows the estimation errors of the upper body posed by the whole-body and the upper-body models, calculated from the data acquired with the D435. Although the estimation errors in the head, right shoulder, and right wrist are higher than those estimated by the whole-body model, as shown in the bold numbers, the overall accuracy of 3D human pose estimation significantly improved by about 17mm when the upper-body model was used to estimate the pose of the upper body ($M = 124.123$, $SD = 30.7033$, $t(8) = 1.94$, $p < .05$).

V. Detection of Deterioration in Sitting Postures

During desk work, people tend to unconsciously adopt a posture that places a strong load on some part of their body. Slouching, a posture where the back is arched and the head is forward, is a typical poor posture when sitting. Slouching causes stiff shoulders and back pain due to the load placed on the neck, shoulders and back to support the weight of the head moving forward. Although the sitter can always be aware of and strive to improve their posture when sitting, it would be difficult for them to maintain such awareness during long hours of desk work.

In this study, we developed a method for detecting deteriorated sitting postures from images acquired from a

Web camera installed above a PC monitor using 3D human pose estimation. The advantage of a webcam is that it is widely used and installed as a standard feature of many laptops, allowing for easy access to our purpose. Furthermore, since there is no physical contact between the sensor device and the user, it is comfortable and does not interfere with movement (Figure 5). As joint angles can be calculated from 3D human pose, the characteristics of those angles in sitting postures can be analyzed.

To assess sitting postures, we focused on the change in the angle formed by the nose, neck, and pelvis. As shown in Figure 6, two different sitting postures change the angles formed by the nose, neck and pelvis. These angles change significantly as the head moves forward. The dotted line shows the empirical threshold for the tolerance of good sitting posture. As the sitting posture deteriorated, the observed angle values would no longer fall within this acceptable range.

VI. Conclusion

In this study, we experimentally constructed a 3D human pose estimation model specialized for a part of the body using a single camera. As a result, we confirmed that the accuracy of 3D human pose estimation by the model specialized for given parts of the body was higher than that by the model for the whole body.

The proposed system to detect posture deterioration during sitting using the 3D human pose model specialized for the upper body confirmed that it can detect postures deterioration based on the changes in angles formed by the nose, neck, and pelvis. In the future, we would like to further improve the reliability of the posture deterioration detection by combining multiple joint angles for practical use.

References

[1] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A Simple Yet Effective Baseline for 3d Human Pose Estimation," arXiv preprint arXiv:1705.03098, pp. 1-10, 2017.

[2] W. Yang et al., "3D Human Pose Estimation in the Wild by Adversarial Learning," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5255-5264, 2018.

[3] O.D.A. Prima et al., "Evaluation of Joint Range of Motion Measured by Vision Cameras," International Journal on Advances in Life Sciences , 11, 3 & 4, pp. 128-137, 2019.

[4] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, 36, 7, pp. 1-15, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification," Proceedings of the IEEE International Conference on Computer Vision, pp. 1026-1034, 2015.

[6] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," Lecture Notes in Computer Science, 9912 LNCS, pp. 483–499, 2016.

 ISBN: 978-1-61208-872-3