A Multiview Depth-based Motion Capture Benchmark Dataset for Human Motion Denoising and Enhancement Research

Nate Lannan, Le Zhou, and Guoliang Fan
School of Electrical and Computer Engineering
Oklahoma State University, Stillwater, OK 74078
(nate.lannan,le.zhou,guoliang.fan)@okstate.edu

Abstract

The field of human motion enhancement is a rapidly expanding field of study in which depth-based motion capture (D-Mocap) is improved to generate a more accurate counterpart for demanding high precision real-world applications. The D-Mocap that is initially generated relies on commercially available SDKs or open source tools to produce the initial skeletal sequence which works best in an ideal front-facing camera setup. This in turn creates a challenging initialization for human motion enhancement when the camera is not positioned in the ideal forward facing position. Currently there are no multiview D-Mocap datasets which have corresponding time-synced and skeleton-matched optical motion capture (Mocap) reference data for view-invariant motion enhancement. We develop a multiview D-Mocap dataset extended from the popular and comprehensive Berkeley MHAD dataset [29]. In addition, we analyze the performance of the D-Mocap data generated through a series of open source tools, highlighting the difficulty and the need to produce robust results in a rear-facing camera setup due to a 21.4% increase in average joint position error over front-facing data. Finally, we analyze the results of some recent human motion enhancement algorithms with regard to a front-facing camera setup versus a rear-facing one.

1. Introduction

Human motion analysis and human pose recognition are well studied fields in computer vision which aim to generate a three-dimensional human model or human joint positions defined in Cartesian space from various sensor technologies (Fig. 1). Human motion data can be highly accurate as with data collected from optical motion capture (Mocap) systems, or less accurate but easily obtained as with data collected from RGB-D (D-Mocap) or inertial sensors. Due to the high complexity and cost of optical Mocap technology, its real-world applications are often limited to a lab setting. On the other hand, low-cost D-Mocap has been applied for some real-world applications, such as gait assessment [8, 15, 28, 33], rehabilitation [34], human mobility analysis [22], and exercise systems [6]. With human motion enhancement and denoising of the low-quality D-Mocap data, depth sensors could become an inexpensive and versatile alternative for clinical applications,

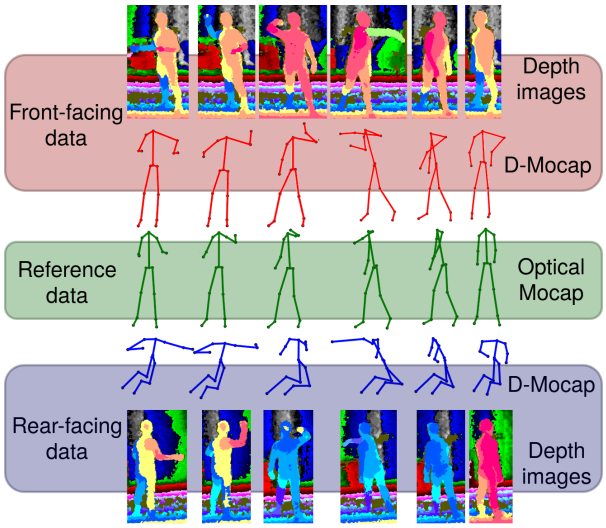

Figure 1. The extended MHAD dataset. Data captured with front and rear facing depth sensors is used to generate D-Mocap data. This data is skeleton matched, time synced, and spatially registered to highly accurate optical Mocap for reference.

Although there are ample datasets that provide clean high quality Mocap data, there are very few that provide low quality D-Mocap motion data with a high quality counterpart time synced for performance evaluation. Most importantly, there are none that we are aware of that provide D-Mocap data taken from multiple angles and time aligned to a highly accurate optical Mocap reference which are desirable for view-robust human motion enhancement.

Without an open-source benchmark dataset, researchers in human motion enhancement have been employing two methods to evaluate their work. In the first method, artificial noise of different-levels or various dropouts are induced to high quality Mocap data to simulate the low-quality D-Mocap data [13, 14, 18–20, 24, 43–45]. This method suffers a significant shortcoming since modern methods like deep learning and nonlinear Kalman filtering have proven very successful at removing induced corruption, while proving less effective at real-world generated motion data [18, 24, 44]. The second approach uses motion data collected in the lab along side high quality optical Mocap in order to compare the improvements made to the lower quality data [13, 18–20, 24, 43–45]. This method reflects a real-world scenario, but is insular in that the motion data is specific to the researcher's task and is not often shared amongst the community. A dataset which provides time-synced and skeleton-matched low and high quality human motion data over a wide range of subjects and actions would provide this emerging field with the means to quantitatively and technically compare competing methodologies.

In addition, as the research of D-Mocap human motion enhancement evolves and progresses, view invariant algorithms will become a necessity and an expectation. One of the major advantages of depth-based human pose estimation is that the subject is not confined by markers and a capture space. To fully take advantage of this boon, an algorithm that improves poor quality rear-view data will be essential. Therefore, a multi-view benchmark dataset which highlights non-optimal depth camera placement will be a critical and a much-needed tool for the advancement of the field of human motion enhancement.

We describe in this work an extension of the Berkeley MHAD dataset [29] which includes low quality D-Mocap data. The MHAD dataset provides a rear-facing depth sensor capture angle for use in testing non-ideal view invariance and can be used to highlight the shortcomings of D-Mocap data capture. MHAD is a widely used dataset spanning fields of study including human motion enhancement [13,19,20,40], human action recognition [11,32], multiview and view invariant action recognition [41], human motion synthesis [14, 31], and human shape reconstruction [16]. MHAD is flexible as well, due to its multimodality, containing data from five different systems (Fig. 2).

The solid foundation provided by the Berkeley MHAD data set, which contains data from 12 subjects performing 11 diverse actions, is an excellent choice to extend to a benchmark D-Mocap data set. It provides a highly accurate time-synchronised optical Mocap reference and multiple depth cameras (Fig. 3) that we have used to create real-world D-Mocap data. Most importantly this dataset includes data from a depth camera placed in a non-ideal rear-facing position. This positioning highlights one of the

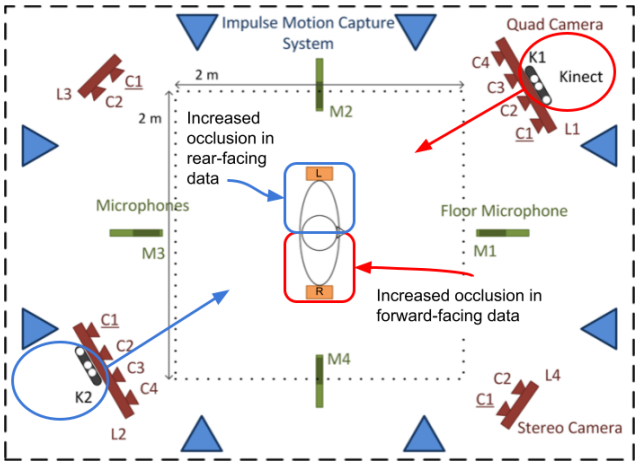

Figure 2. The MHAD recording layout [29]. The capture angles of the depth sensors produces higher error in the upper portion of the body that is self-occluded due to the off-axis sensor position.

major weaknesses of D-Mocap data generation algorithms and will be a critical tool in human motion enhancement research as the community tackles the crucial problem of view invariance. This benchmark dataset is publicly available[1] so that researchers may freely compare results on the same D-Mocap data, something that is lacking in the human motion enhancement community.

Figure 3. The MHAD dataset provides two depth sensor capture positions, one forward-facing and the other rear-facing [29].

2. Related Work

In this section we will highlight several current methodologies for enhancement of D-Mocap. We will concentrate on research that improves D-Mocap skeleton data and not research that seeks to improve human motion estimation from depth images.

[1] http://vcipl-okstate.org/pbvs/bench/

Table 1. Overview of some of the state-of-the-art in human motion enhancement systems.

	Methodology		Run Mode			Sensor configuration				Motion data type						Metrics		
	Filtering based	Learning based	Real-time	On-line	Off-line	Single sensor	Multiple sensors	Single subjective	Multiple subjective	Partial joints	Full body	Single activity	Multiple activities	single view	Multiple views	Joint positions	Bone lengths	Joint angles
EKF [35]	√		√			√		√		√		√		√		√		
UKF [21]	√		√			√		√		√		√		√		√		
TKF [26, 44]	√		√			√			√		√		√	√		√	√	
TKF-DE [44]	√				√	√			√		√		√	√		√	√	√
$Autoencoder$ [13]		√			√	√			√		√		√	√		√	√	√
KF-$guided\ Autoencoder$ [18]		√			√	√			√		√		√	√		√		
BRA-P [24]		√			√	√			√		√		√	√		√		

2.1. Motion Filtering Methods

The intent of traditional filtering methods in human motion enhancement is to remove noise in joint position data, thereby arriving at a more accurate result. The most popular and effective of the filtering methods is the Kalman filter (KF) [17]. Researchers have used the linear Kalman filter to constrain joint dynamics with the objective of keeping bone length constant [39], but most commonly researchers have gravitated toward using a nonlinear KF due to the nonlinear nature of D-Mocap data. The most common of the nonlinear Kalman filters are the extended Kalman filter (EKF) and unscented Kalman filter (UKF) [21, 35, 36]. Recently, researchers have begun to explore the merits of the Tobit Kalman filter since it is particularly well suited to deal with the nonlinearity and non-Gaussian nature of D-Mocap data due to the censored nature of occluded D-Mocap [1, 26, 27, 44, 45]. Self occlusion errors are significantly reduced through the use of the Tobit model [26], especially when adapting the censor limits of the TKF [27, 44, 45].

The introduction of genetic algorithms (GAs) into filter systems like the TKF and particle filters, have shown to be successful in enhancing D-Mocap as well [7, 38, 44, 45]. A multi-objective GA is used in conjunction with a particle filter in [38]. The GA is used to constrain bone length while the data is being filtered with the particle filter. Much the same technique is used in [7] to constrain bone length with a differential evolutionary algorithm (DE). DE is used to restrict bone length while employing the TKF in [44], and is combined with a particle filter initialized with the Tobit model in [43].

2.2. Motion Learning Methods

Of the data driven learning methods currently exploring human motion enhancement, the most popular methods are dimensionality reduction, sparse coding, Gaussian Process (GP) models, and deep learning. The Greedy Kernel PCA [10] is used for dimensionality reduction in [37], and thereby represent human motion in the Hilbert space. As is common with dimensionality reduction, this is done to remove any characteristics of the D-Mocap data that are aberrant from human motion. Researchers in [9, 42] use sparse coding dictionaries to in conjunction with bone length and smoothing models to diminish noise and outliers in D-Mocap. The works [6, 25] use GP to constrain velocity variation in optimization, and to map D-Mocap to optical Mocap respectively.

Not surprisingly, the most often used machine learning method in recent human motion enhancement research is deep learning. With the rise in parallel processing power has come a surge in powerful deep learning techniques for many optimization problems, and human motion enhancement is no exception. The work in [30] uses two interconnected recurrent neural networks. One network is trained on the joint positions of the human motion, and the other is trained on the joint velocities, thereby learning two aspects of the natural kinematics of the human motion. Most of the deep learning solutions in recent literature deal with learning a lower dimensional representation of human motion, thereby extracting noise from data as it is mapped to the lower dimension [3, 13, 14, 18–20, 23, 24, 40]. In [3] the authors use three types of temporal encoders in an attempt to eliminate data that is disparate from natural human motion. Similarly, authors in [40] break up a network into 3 parts, a temporal section, a spacial section, and a residual section. The temporal section is a bidirectional long short term memory (LSTM) encoder that learns time dependencies in the motion data. The spacial and residual sections are both fully connected; the former learns spacial interdependencies of joint positions and the latter learns to remove high frequency noise from the data.

The work in [13, 14] uses a convolutional autoencoder which simultaneously learns temporal and positional interdependencies of data by training on a large set of varied Mocap data and learning a motion manifold for valid human motion. However, through this method, some of the original kinematics of the human motion are lost as the motion manifold does not take into account the kinematics of the test data. Researchers in [23, 24] have addressed this problem

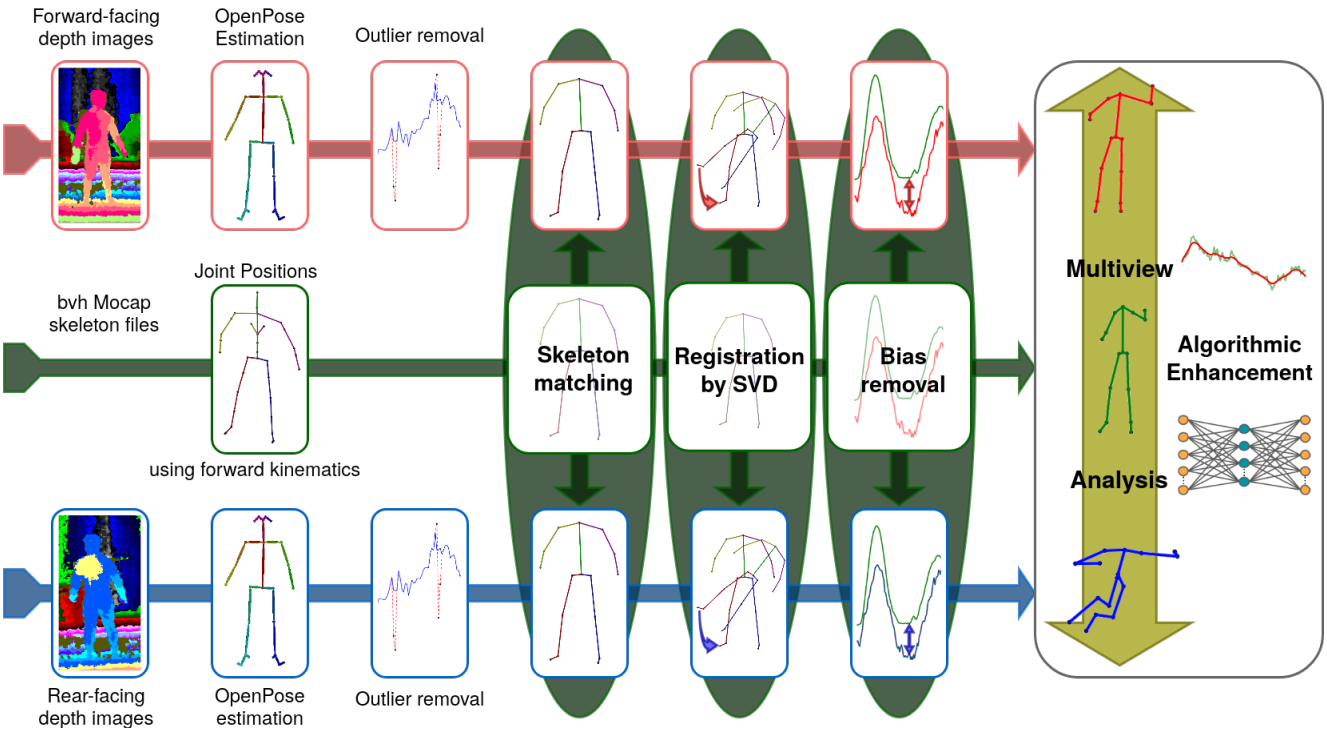

Figure 4. Extending the MHAD dataset to include D-Mocap data. D-Mocap is first generated from depth images while joint positions for the optical Mocap reference are are calculated through forward kinematics. Outliers are then removed from the D-Mocap and the D-Mocap and Mocap reference are matched to the same skeletal model. Finally, the D-Mocap is registered to the Mocap reference through SVD, and bias is removed to produce a dataset that can be used for multiview analysis and algorithmic enhancement.

by incorporating a bidirectional LSTM in an autoencoder to learn time interdependencies of joints, thereby preserving some of the data's natural kinematics. Similarly [19,20] uses a set of target motion filtered with a TKF to optimize the output of the autoencoder in the autoencoder's latent space. This method preserves the natural kinematics of the motion while still adhering to the manifold learned by the autoencoder.

We provide a brief overview of some of the tactics mentioned in this section, including method categories, run modes, data collecting, and outcomes analysis. The details are presented in Table 1. None of these recent methods have attempted to tackle enhancement of multiview D-Mocap or D-Mocap generated from rear facing depth sensors. This is because available software tools that are used to generate D-Mocap struggle to estimate human pose from capture angles that are not forward facing. In order to explore the challenge of enhancing this sub-par data, we need a common benchmark dataset that uses real-world D-Mocap, provides multiview with a rear-facing camera, and is temporally synced and spatially registered to a highly accurate optical Mocap reference. Our work in providing the MHAD dataset for multiview D-Mocap provides this benchmark to the community.

3. Dataset Generation

3.1. Skeleton Generation

The purpose of this work is to extend the MHAD dataset to include real-world D-Mocap data. This process is illustrated in Figure 4. When the RGB-D images of the MHAD dataset are processed with the *ros_openpose_rgbd* [5] program, low-quality D-Mocap data is generated. This procedure can be divided into two steps. To begin, OpenPose [4] estimates the locations of joints in two-dimensional space from the appropriate pixel position in the frame's RGB image. Second, the joint positions are placed in three-dimensional space using depth image data and the initial two-dimensional values. Although this method generates estimates of joint motion trajectories similar to those generated by publicly available SDKs, ocasionally some extreme joint positions are created in the process resulting in outliers. Due to the desire to generate a dataset that closely represents data captured using SDK software, we processed the obtained data through a Hample Filter to reduce the effect of outliers. In Hampel filtering, the median joint position value was calculated using a seven-frame window. This median value can then be used to estimate standard deviation using median absolute deviation (MAD), by the equa-

tion $\sigma = 1.4826$(MAD). If the value of the joint position is more than three standard deviations from the mean, the value is substituted with the median of the window (Fig. 5).

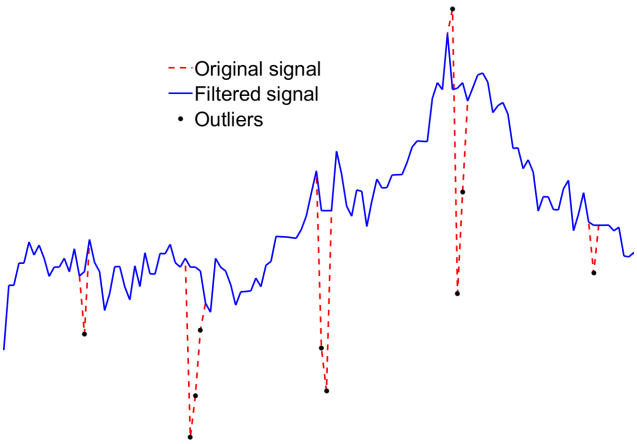

Figure 5. Outliers are removed from D-Mocap using Hampel filtering. This process is done to ensure data error is akin to commercially available D-Mocap SDKs like Nuitrack and Kinect.

3.2. Skeleton Matching

The default skeletal model used in *ros_openpose_rgbd* is *keypoints_pose_COCO_18* which contains 18 body joints (Fig. 6). This model features a split human torso which is not common in most human motion enhancement work and is not easily comparable to the Mocap data from MHAD. In comparison, the *keypoints_pose_BODY_25* model provided by OpenPose contains 16 joints that can be matched with Mocap skeleton data generated from the MHAD dataset (Fig. 6). The MHAD optical Mocap motion data is converted from motion data recorded from the Mocap system and placed in the global frame using forward kinematics on the bvh files. This results in a 35 joint model(Fig. 7 (a)). The MHAD Mocap model is matched with the D-Mocap *BODY_25* model to generate a skeleton model of 16 common joints that can be compared across both the Mocap and D-Mocap data (Fig. 7 (b)). In contrast, in the *COCO* model, there are only 13 joint positions common with the optical Mocap model. These 13 joints mostly consist of joint positions central to the human body, and some extremities such as toe joints are lost. These missing joints are the most suceptable to errors from occlusion and therefore contribute more to the evaluation of the human motion enhancement method's efficacy.

3.3. Data Registration

The sets of Mocap and D-Mocap data are recorded in their own respective 3D spaces and must be matched so that

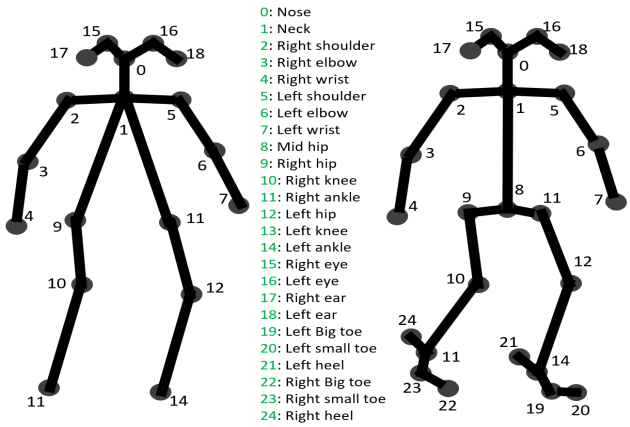

Figure 6. Two skeleton models included in OpenPose, on the left is COCO_18 and on the right is BODY_25. BODY_25 was chosen for skeleton matching since it had more joints in common with the optical Mocap skeleton model provided by MHAD.

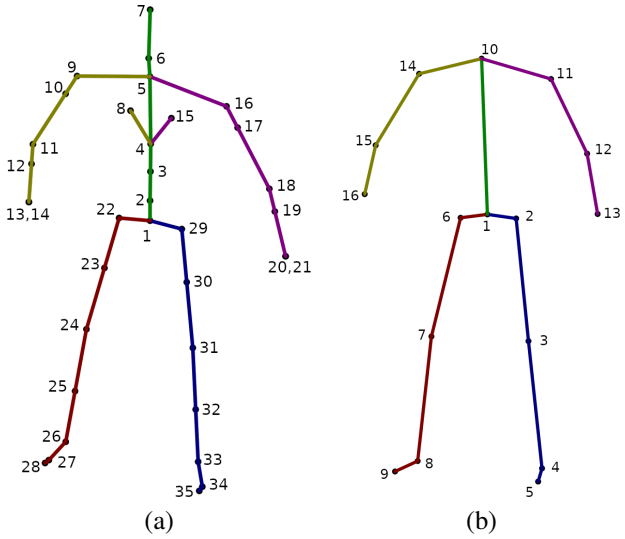

Figure 7. The optical Mocap skeletal model with 35 joints provided by MHAD is depicted on the left. The final skeleton model used in the Extended MHAD Dataset is shown on the right. These are the 16 joints that are common between the BODY_25 model and the optical Mocap skeleton model.

a valid comparison may be made. To calculate the rigid transform between the two skeletons, we use singular value decomposition (SVD) [2]. In this method, a centroid for each entire sequence of data is calculated using,

$$\mathbf{C} = \frac{1}{N \cdot I} \sum_{i=1}^{I} \sum_{j=1}^{N} p_{i,j}, \tag{1}$$

where N is the total number of joints in the skeletal model I is the number of frames in the sequence, and $p_{i,j}$ is a particular joint position for a particular frame. Using 1, an

intermediate matrix $\mathbf{H}$ is calculated from the D-Mocap data $\overline{\mathbf{X}}$, Mocap data $\mathbf{X}$, and the centroids of each dataset ($\mathbf{C}_{\overline{\mathbf{X}}}$, $\mathbf{C}_{\mathbf{X}}$),

$$\mathbf{H} = (\overline{\mathbf{X}} - \mathbf{C}_{\overline{\mathbf{X}}}) \times (\mathbf{X} - \mathbf{C}_{\mathbf{X}})^T. \quad (2)$$

Next, using SVD on $\mathbf{H}$ from 2, the factorization of $\mathbf{H}$ produces $\mathbf{U}$ and $\mathbf{V}$. The benefit of doing this is that these matrices can be used for registration by calculating the necessary rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ by the following:

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = SVD(\mathbf{H}), \quad (3)$$

$$\mathbf{R} = \mathbf{U} \times \mathbf{V}^T \quad (4)$$

$$\mathbf{t} = \mathbf{C}_{\mathbf{X}} - \mathbf{R} \times \mathbf{C}_{\overline{\mathbf{X}}}. \quad (5)$$

Once $\mathbf{R}$ and $\mathbf{t}$ are obtained, a rigid body transform on the D-Mocap aligns the data as closely as possible in the same 3-D space with the Mocap reference,

$$\widetilde{\mathbf{X}} = \mathbf{R} \times \overline{\mathbf{X}} + \mathbf{t}, \quad (6)$$

where $\widetilde{\mathbf{X}}$ is the point registered version of the D-Mocap data.

3.4. Bias Removal

The skeletal structure calculated by the optical Mocap system and *ros_openpose_rgbd* define the locations of each joint slightly different from one another even though they are estimating the same joint. This results in a constant offset which is not due to noise or occlusion, but is instead just a difference in definition. Any analysis done on the pure data of these two data sets would be subject to this bias and may yield erroneous results. In order to correct for this difference in joint definition, a bias is calculated between the D-Mocap data and the Mocap data. This is done by calculating a Euclidian distance for each joint j in each frame i, between the D-Mocap data and Mocap data. These values are then averaged over all frames using:

$$Bias_j = \frac{1}{I} \sum_{i=1}^{I} \widetilde{\mathbf{X}}_{i,j} - \mathbf{X}_{i,j}, \quad (7)$$

where $\widetilde{\mathbf{X}}$ is a matrix of D-Mocap data, $\mathbf{X}$ is a matrix of Mocap, and I is the total number of frames. The bias is calculated for each joint so that it may be removed from the D-Mocap data. This yeilds a 0 mean error between the D-Mocap and optical Mocap for each joint position.

4. Multiview analysis

4.1. Multiview Comparison of D-Mocap

One clear problem of using commercially available and open pose tools for D-Mocap generation is the weakness that these methods have with rear facing depth cameras. Figure 1 highlights this problem over a series of corresponding frames of D-Mocap data generated from the front-facing camera and the rear-facing camera of the extended MHAD dataset. We can see that the *ros_openpose_rgbd* opensource tool struggles to estimate the same subject when captured from the rear. This is why view invariance in human motion enhancement is so important and why the community needs a dataset that accurately represents the problem with D-Mocap captured from the rear.

To numerically express this problem of rear-facing D-Mocap, we have done a comparison between the D-Mocap data generated from the front-facing depth camera and the rear-facing counterpart. This comparison was done by joint position error with respect to the optical Mocap reference. This calculation is given in the following equation,

$$meanErr_j = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{X}_{i,j} - \widehat{\mathbf{X}_{i,j}} \right\|_2, \quad (8)$$

where $meanErr_j$ is the average distance the enhanced D-Mocap motion data, $\widehat{\mathbf{X}}$, is from Mocap motion data over all frames, i, for a particular joint j.

In comparing the D-Mocap columns in Table 2, we see how the algorithm used to generate human motion data struggles when using depth images from the rear facing camera. This problem is quantified by a 21.4% increase in average joint position error, a $1.8cm$ increase from the results from the front-facing camera. We notice, the few joints that had better results in the rear-facing dataset are generally internal to the body structure and do not differ in performance of the forward-facing dataset nearly as much as the peripheral joints like the hands and feet. We do notice one exception, the joints on the upper right quadrant of the body perform worse in the front-facing data than the rear-facing data. The reason for this is that the Kinect cameras are angled to the human subject rather than directly facing the front and rear of the human subject (Fig. 2). Due to these placements, there is more occlusion of the joints in the right half of the body in the front-facing data, and in the left half of the body for the rear-facing data. We also see that the worst result of both cameras is joint 13, the left hand, which is susceptible to inaccuracies from capturing from the rear-view as well as suffering from a higher rate of occlusion. For convenience, joint numbers in Table 2 on the left side of the body are indicated with an (L) and joints on the right side of the body are indicated with an (R).

Fortunately, the increase in error due to occlusion on each side of the body is actually a benefit to researchers who want to study joints that suffer from higher rates of occlusion than others. The rear placement of the depth camera is obviously a problem in current D-Mocap generation methods, but with a dataset that makes multiview real-world D-Mocap data readily available, this problem can be attacked as a community, using a standard set of motion data.

Table 2. Front and Rear Facing MHAD Motion Enhancement Results in (cm). (L) and (R) depict left or right side of the body.

	Front Facing					Rear Facing				
Joints	D-Mocap Data	EKF [35] Enhanced	UKF [21] Enhanced	TKF [44] Enhanced	Autoencoder [13] Enhanced	D-Mocap Data	EKF [35] Enhanced	UKF [21] Enhanced	TKF [44] Enhanced	Autoencoder [13] Enhanced
1	1.5	3.2	2.4	2.0	2.2	2.0	3.1	2.5	1.8	2.9
2 (L)	3.6	4.1	2.6	2.9	2.4	2.2	3.5	2.9	1.8	3.3
3 (L)	3.4	4.6	3.1	2.9	4.0	3.1	3.6	3.1	2.4	4.1
4 (L)	3.7	4.2	3.7	3.0	4.7	5.2	4.7	4.4	3.8	5.3
5 (L)	6.7	5.6	4.0	4.2	7.5	8.8	7.3	6.9	5.9	7.0
6 (R)	3.6	4.5	2.6	2.9	2.4	5.1	5.4	4.8	3.4	3.6
7 (R)	4.0	4.1	3.3	3.0	4.4	24.4	17.0	16.5	14.2	13.9
8 (R)	4.2	4.7	3.9	3.2	5.0	5.2	4.6	4.2	3.2	5.0
9 (R)	8.7	5.6	4.0	5.3	7.5	18.0	11.9	11.6	10.1	10.3
10	5.2	5.6	6.1	4.3	4.0	2.6	4.6	3.7	2.1	4.4
11 (L)	5.2	6.4	7.9	4.4	4.2	3.4	4.9	4.1	2.6	4.8
12 (L)	8.2	7.8	8.0	6.7	6.2	15.3	13.7	13.0	10.5	12.0
13 (L)	19.4	13.6	13.3	11.6	13.2	39.6	30.4	28.8	27.1	26.6
14 (R)	10.7	6.2	7.0	6.3	6.0	2.7	4.7	3.5	2.1	4.0
15 (R)	16.1	11.0	10.9	9.9	11.7	7.6	6.9	7.3	5.2	8.7
16 (R)	30.4	17.0	17.0	16.3	19.5	17.9	16.3	14.9	11.0	15.1
Ave.	8.4	6.7	6.2	5.5	6.6	10.2	8.9	8.3	6.7	8.2

4.2. Multiview Human Motion Enhancement

We used four methods of human motion enhancement to illustrate the applicability of our extended MHAD dataset. These methods include three nonlinear Kalman filtering methods and one deep learning method, a convolutional autoencoder first proposed by Holden *et al*. [13, 14]. The publicly available autoencoder [12] was retrained using the 16 joint skeletal structure described in Section 3.2. The convolutional autoencoder was trained on a homogeneous skeleton where multiple subjects were all retargeted to ensure all bone lengths and skeletal sizes remained the same. However, in testing we opted to only scale the D-Mocap as we are attempting to analyze the dataset more than the human motion enhancement method.

The modification of the D-Mocap for the autoencoder was a two step process. First, using subject T-poses each skeleton was scaled to match the size of the training skeleton with regard to the distance of the subject's hips to the floor. Second, after the data was recovered from the autoencoder the resultant human motion was re-scaled using the inverse of the scaling value in step one and any constant bias between the recovered skeleton and the MHAD Mocap data was removed using 7.

The joint positions of the enhanced human motion data are then evaluated on a joint by joint basis using the euclidean distance from the MHAD Mocap data averaged over all frames (8). These joint-by-joint values are given in Table 2 for the original D-Mocap data and all four enhancement methods. In addition, a qualitative analysis of each enhancement method over corresponding frames of human motion data is shown in Figure 8. This is an example of how the extended MHAD dataset can be used to compare various methods of human motion enhancement in a multiview manner.

5. Conclusion

Human motion enhancement of D-Mocap makes low-cost depth sensors promising and affordable devices for real-world clinical and health-related applications. However, this research is now at a crossroads. In order to exploit the markerless and mobile capabilities of depth sensors, we must tackle view invariance and we also need a public benchmark dataset for algorithm evaluation. This subject has largely been ignored by human motion enhancement researchers for two reasons. First, the software tools used to generate D-Mocap do a poor job when the depth sensor is placed in a non-forward facing position as shown in Figures 1, 8, and Table 2. Second, the community lacks a real-world dataset that provides multiview D-Mocap that accentuates both ideal front-facing data and non-ideal rear-facing data. We have provided a publicly available dataset that extends MHAD to include D-Mocap that has been spatially registered and temporally synced to highly accurate optical Mocap data for reference. This dataset provides the necessary tools for the human motion enhancement community to begin tackling the difficult problem of view invariance and especially data generated from a rear-facing depth sensor.

6. Acknowledgment

This work is supported in part by the US National Institutes of Health (NIH) Grant R15 AG061833 and the Oklahoma Center for the Advancement of Science and Technology (OCAST) Health Research Grant HR18-069.

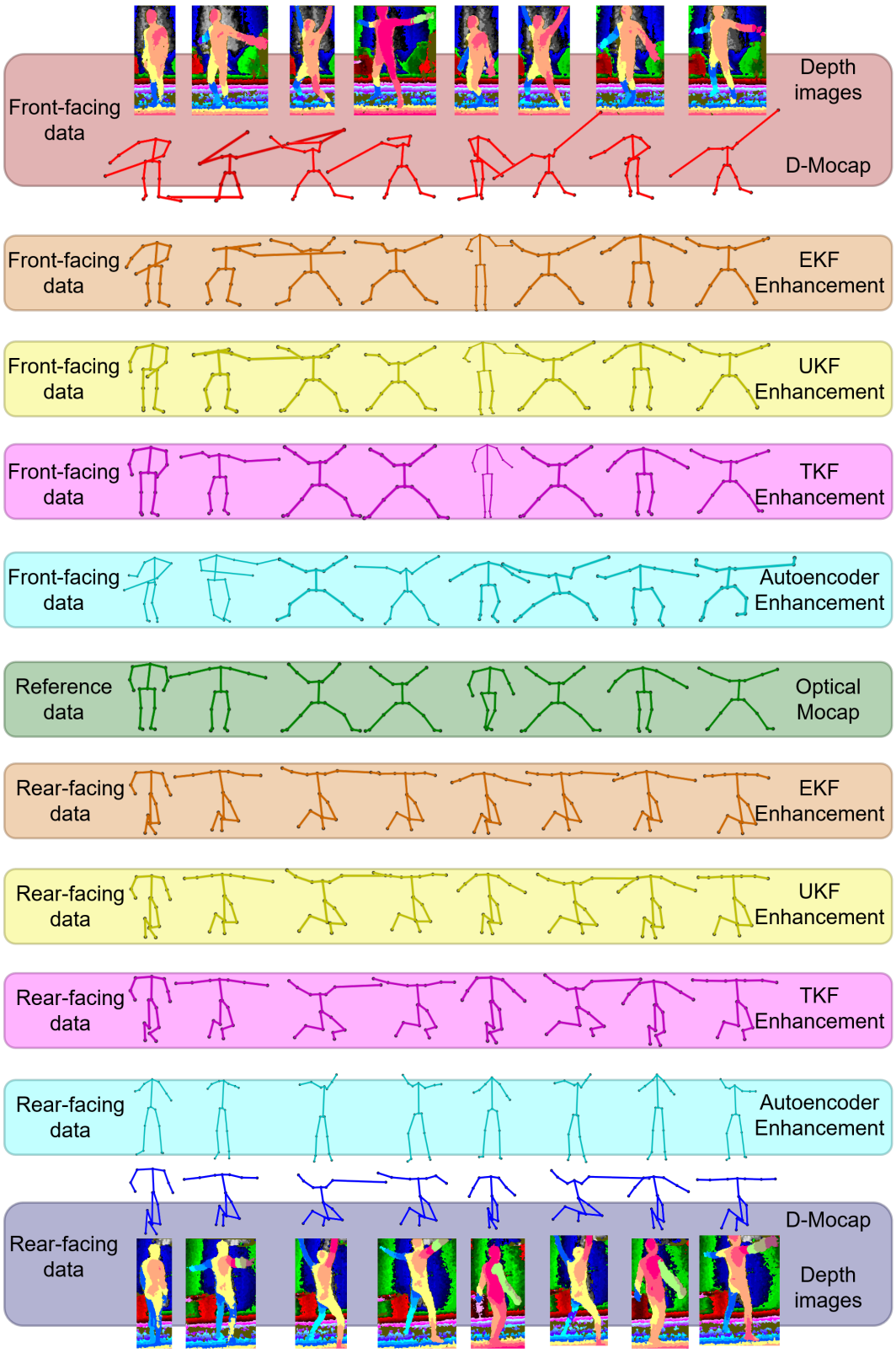

Figure 8. Qualitative analysis of the extended MHAD dataset with 4 benchmark enhancement methods applied to both the front-facing and rear-facing data.

References

[1] B. Allik, C. Miller, M. J. Piovoso, and R. Zurakowski. Nonlinear estimators for censored data: A comparison of the ekf, the ukf and the tobit kalman filter. In *2015 American Control Conference (ACC)*, pages 5146–5151, 2015. 3

[2] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-D point sets. *IEEE Trans. PAMI*, 9(5):698–700, 1987. 5

[3] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proc. IEEE CVPR*, 2017. 3

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. PAMI*, 43(1):172–186, 2019. 4

[5] F. Chen. ros_openpose_rgbd, 2019. Code from https://github.com/felixchenfy/ros_openpose_rgbd. 4

[6] A. T. Chiang, Q. Chen, Y. Wang, and M. R. Fu. Kinect-based in-home exercise system for lymphatic health and lymphedema intervention. *IEEE Journal of Translational Engineering in Health and Medicine*, 6:1–13, 2018. 1, 3

[7] P. Das, K. Chakravarty, D. Chatterjee, and A. Sinha. Improvement in Kinect based measurements using anthropometric constraints for rehabilitation. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, 2017. 3

[8] Robert J Dawe, Lei Yu, Sue E Leurgans, Timothy Truty, Thomas Curran, Jeffrey M Hausdorff, Markus A Wimmer, Joel A Block, David A Bennett, and Aron S Buchman. Expanding instrumented gait testing in the community setting: A portable, depth-sensing camera captures joint motion in older adults. *PloS one*, 14(5):e0215995, 2019. 1

[9] Y. Feng, M. Ji, J. Xiao, X. Yang, J. J Zhang, Y. Zhuang, and X. Li. Mining spatial-temporal patterns and structural sparsity for human motion data denoising. *IEEE Trans. Cybernetics*, 45(12):2693–2706, 2014. 3

[10] V. Franc and V. Hlaváč. Greedy kernel principal component analysis. In *Cognitive Vision Systems*, pages 87–105. Springer, 2006. 3

[11] N. Golestani and M. Moghaddam. Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks. *Nature communications*, 11(1):1–11, 2020. 2

[12] Daniel Holden. The orange duck, 2019. Data retrieved from theorangeduck.com, http://theorangeduck.com/. 7

[13] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4):138:1–138:11, July 2016. 2, 3, 7

[14] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, SA '15, pages 18:1–18:4, New York, NY, USA, 2015. ACM. 2, 3, 7

[15] Andrew Hynes, Stephen Czarnuch, Megan Kirkland, and Michelle Ploughman. Spatiotemporal gait measurement with a side-view depth sensor using human joint proposals. *IEEE Journal of Biomedical and Health Informatics*, 2020. 1

[16] H. Jiang, J. Cai, and J. Zheng. Skeleton-aware 3d human shape reconstruction from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5431–5441, 2019. 2

[17] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960. 3

[18] Nate Lannan and Guoliang Fan. Filter guided manifold optimization in the autoencoder latent space. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 3

[19] Nate Lannan, Le Zhou, and Guoliang Fan. Human motion enhancement via tobit kalman filter-assisted autoencoder. *IEEE Access*, 10:29233–29251, 2022. 2, 3, 4

[20] N. Lannan, L. Zhou, G. Fan, and J. Hausselle. Human motion enhancement using nonlinear kalman filter assisted convolutional autoencoders. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 1008–1015, Los Alamitos, CA, USA, oct 2020. IEEE Computer Society. 2, 3, 4

[21] Anders Boesen Lindbo Larsen, Søren Hauberg, and Kim Steenstrup Pedersen. Unscented Kalman filtering for articulated human tracking. In Anders Heyden and Fredrik Kahl, editors, *Image Analysis*, pages 228–237, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. 3, 7

[22] Daniel Leightley, Jamie S McPhee, and Moi Hoon Yap. Automated analysis and quantification of human mobility using a depth sensor. *IEEE journal of biomedical and health informatics*, 21(4):939–948, 2016. 1

[23] S. Li, Y. Zhou, H. Zhu, W. Xie, Y. Zhao, and X. Liu. Bidirectional recurrent autoencoder for 3d skeleton motion data refinement. *Computers & Graphics*, 81:92–103, 2019. 3

[24] S. Li, H. Zhu, L. Zheng, and L. Li. A perceptual-based noise-agnostic 3d skeleton motion data refinement network. *IEEE Access*, 8:52927–52940, 2020. 2, 3

[25] Z. Liu, L. Zhou, H. Leung, and H. P. H. Shum. Kinect posture reconstruction based on a local mixture of gaussian process models. *IEEE Transactions on Visualization and Computer Graphics*, 22(11):2437–2450, 2016. 3

[26] K. Loumponias, Nicholas Vretos, Petros Daras, and G. Tsaklidis. Using Kalman filter and Tobit Kalman filter in order to improve the motion recorded by Kinect sensor II. In *Proceedings of the 29th Panhellenic Statistics Conference*, 2016. 3

[27] K. Loumponias, N. Vretos, G. Tsaklidis, and P. Daras. An improved Tobit Kalman filter with adaptive censoring limits. *Circuits Syst Signal Process*, 39:5588–5617, 2020. 3

[28] Björn Müller, Winfried Ilg, Martin A Giese, and Nicolas Ludolph. Validation of enhanced kinect sensor based motion capturing for gait assessment. *PloS one*, 12(4):e0175813, 2017. 1

[29] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley MHAD: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60. IEEE, 2013. 1, 2

[30] Youngbin Park, Sungphill Moon, and Il Hong Suh. Tracking human-like natural motion using deep recurrent neural networks. *CoRR*, abs/1604.04528, 2016. 3

[31] D. Pavllo, C. Feichtenhofer, M. Auli, and D. Grangier. Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision*, pages 1–18, 2019. 2

[32] Y. Qin, L. Mo, C. Li, and J. Luo. Skeleton-based action recognition by part-aware graph convolutional networks. *The visual computer*, 36(3):621–631, 2020. 2

[33] Ana Patrícia Rocha, Hugo Miguel Pereira Choupina, Maria do Carmo Vilas-Boas, José Maria Fernandes, and João Paulo Silva Cunha. System for automatic gait analysis based on a single rgb-d camera. *PloS one*, 13(8):e0201728, 2018. 1

[34] Joe Sarsfield, David Brown, Nasser Sherkat, Caroline Langensiepen, James Lewis, Mohammad Taheri, Christopher McCollin, Cleveland Barnett, Louise Selwood, Penny Standen, et al. Clinical assessment of depth sensor based pose estimation algorithms for technology supervised rehabilitation applications. *International journal of medical informatics*, 121:30–38, 2019. 1

[35] Jody Shu, Fumio Hamano, and John Angus. Application of extended kalman filter for improving the accuracy and smoothness of kinect skeleton-joint estimates. *Journal of Engineering Mathematics*, 88(1):161–175, 2014. 3, 7

[36] Seyoon Tak and Hyeong-Seok Ko. A physically-based motion retargeting filter. *ACM Trans. Graph.*, 24(1):98–117, Jan. 2005. 3

[37] T. Tangkuampien and D. Suter. Human motion de-noising via greedy kernel principal component analysis filtering. In *Proc. ICPR*, 2006. 3

[38] S. R. Tripathy, K. Chakravarty, and A. Sinha. Constrained particle filter for improving Kinect based measurements. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 306–310, 2018. 3

[39] S. R. Tripathy, K. Chakravarty, A. Sinha, D. Chatterjee, and S. K. Saha. Constrained Kalman filter for improving Kinect based measurements. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4, 2017. 3

[40] H. Wang, E. S. L. Ho, H. P. H. Shum, and Z. Zhu. Spatio-temporal manifold learning for human motions via long-horizon modeling. *IEEE Transactions on Visualization and Computer Graphics*, 27(1):216–227, 2021. 2, 3

[41] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[42] J. Xiao, Y. Feng, M. Ji, X. Yang, J. J. Zhang, and Y. Zhuang. Sparse motion bases selection for human motion denoising. *Signal Processing*, 110:108–122, 2015. 3

[43] Le Zhou, Nate Lannan, and Guoliang Fan. Joint optimization of kinematics and anthropometrics for human motion denoising. *IEEE Sensors Journal*, 22(5):4386–4399, 2022. 2, 3

[44] L. Zhou, N. Lannan, G. Fan, and J. Hausselle. A hybrid approach to human motion enhancement under kinematic and anthropometric constraints. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 1028–1035, Los Alamitos, CA, USA, oct 2020. IEEE Computer Society. 2, 3, 7

[45] L. Zhou, N. Lannan, G. Fan, and J. Hausselle. Human motion enhancement via joint optimization of kinematic and anthropometric constraints. *EAI Endorsed Transactions on Bioengineering and Bioinformatics*, 1:e1, 2021. 2, 3