3D HUMAN POSE ESTIMATION FROM MULTI-VIEW RGB IMAGES

by

Hüseyin Temiz

B.S., Computer Engineering, Boğaziçi University, 2011

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering
Boğaziçi University
2019

3D HUMAN POSE ESTIMATION FROM MULTI-VIEW RGB IMAGES

APPROVED BY:

Prof. Lale Akarun (Thesis Supervisor)	.....
Assist. Prof. Berk Gökberk (Thesis Co-supervisor)	.....
Assist. Prof. F. Başak Aydemir	.....
Assist. Prof. M. Furkan Kıraç	.....

DATE OF APPROVAL: 24.07.2019

ACKNOWLEDGEMENTS

I would like to dedicate this thesis to my beloved wife Yasemin, my father Yusuf, my mother Gülderen, and my sister Gül for their unconditional love and support, even sometimes there were kilometers between us.

In the latest years of my undergraduate period, İsmail Arı has been a great inspiration for me to start my academic journey. His motivation in science will always guide me. I want to offer my gratitude to him.

I would like to thank my previous advisor Prof. Cem Ersoy for supporting me during my M.S. studies. He has spent a large amount of time to mentor me and has believed in me for a long time.

I would like to offer my deep gratitude to my instructors and thesis supervisors Prof. Lale Akarun and Assist. Prof. Berk Gökberk, who have given me a second chance to continue my academic career. I have been able to finish my M.S. thesis thanks to this opportunity and their precious supervision. Their constructive criticisms and encouragements made me feel I have led me to learn a lot.

I would like to thank all my colleagues at the Computer Center, especially Mutlu Tunç, Berk Gülenler and Şener Ataş for their savvy manners and providing a suitable environment to handle harsh times.

I especially would like to offer my sincere gratitude to my friends and colleagues. Müjde Aktaş, for her contributions and planful guidance. Ecem Yürük and Faruk Açan, for their critical helps and being awesome old friends.

Finally, I would like to thank my thesis committee; Assist. Prof. M. Furkan Kıraç and Assist. Prof. F. Başak Aydemir for their precious feedback.

ABSTRACT

3D HUMAN POSE ESTIMATION FROM MULTI-VIEW RGB IMAGES

Recovery of a 3D human pose from cameras has been the subject of intensive research in the last decade. Algorithms that can estimate the 3D pose from a single image have been developed. At the same time, many camera environments have an array of cameras. In this thesis, after aligning the poses obtained from single-view images using Procrustes Analysis, median filtering is utilized to eliminate outliers to find final reconstructed 3D body joint coordinates. Experiments performed on the CMU Panoptic, MPI_INF_3DHP, and Human3.6M datasets demonstrate that the proposed system achieves accurate 3D body joint reconstructions. Additionally, we observe that camera selection is useful to decrease the system complexity while attaining the same level of reconstruction performance. We also derive that dynamic camera selection has a more significant impact on reconstruction accuracy as against static camera selection.

ÖZET

ÇOK AÇILI GÖRÜNTÜLERDEN 3 BOYUTLU İNSAN POZU ÇIKARIMI

Son 10 yılda, görüntülerden 3 boyutlu insan pozu çıkarımı yoğun araştırma konularından biri. Tek bir görüntüden 3 boyutlu poz çıkaran algoritmalar geliştirildiler. Bununla beraber, çok fazla kameranın olduğu kurulumlar da mevcut. Bu tezde, Procrustes Analiz tekniğini kullanarak tek görüntüden elde edilmiş pozları hizaladıktan sonra aykırı değerlerden kurtulup nihai 3 boyutlu pozun kritik noktalarının kordinatlarını bulabilmek için medyan filtreleme kullanacağız. CMU Panoptic, MPI_INF_3DHP ve Human3.6M verisetlerinde yaptığımız deneyler önerdiğimiz sistemin insan bedenindeki kritik noktaları birleştirmesini hassas bir şekilde başarıyor. Ayrıca, kamera seçiminin, birleştirme performansını koruyarak sistem karmaşıklığını düşürmede faydalı olduğunu gözlemledik. Dinamik kamera seçiminin statik kamera seçime kıyasla birleştirme başarımı üzerinde belirgin bir etkisi olduğununa da ulaştık.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

s	Scaling factor
R	Rotation matrix
t	Translation vector
J	Number of joints
y_i	3D location of joint i in target pose
x_i	3D location of joint i in pose to be aligned
n	Usable camera count in scene
m	Selected camera count
cam_i	Camera i
I_i	Image of camera i
$pose_i$	Predicted pose from I_i
$avgPose_t$	Average pose for time t
$procrustPose_i$	Procustes aligned $pose_i$ to $avgPose_t$
GT_j	3D ground truth location of joint j
$alignedPose_j$	3D location of joint i in aligned pose

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
4CAMs	4 Cameras
3CAMs	3 Cameras
2CAMs	2 Cameras
RGB	Red Greeen Blue
MPJPE	Mean Per Joint Position Error
PA-MPJPE	Procrustes Aligned Mean Per Joint Position Error
HD	High Definition
VGA	Video Graphics Array
SMPL	Skinned Multi-Person Linear Model
DNN	Deep Neural Network
CMU	Carnegie Mellon University
HMR	Human Mesh Recovery
GT	Ground Truth
ConvNet	Convolutional Neural Network
mocap	Motion Capture
PPS	Predictions Per Second

1. INTRODUCTION

The human communication system with each other and the environment is very complicated. Body pose involves much information about behavior, willing, and our current activity. Hence, the developed methods about body pose estimation can be employed in many daily applications.

Practical applications of human pose information are excessive. For example, an autonomous car can predict the intention of pedestrians by analyzing their body. Besides, older people or people with disabilities need assistance in their daily tasks and routines. Robots can utilize the body pose to fulfill the need of assistance.

Estimating a person's pose in 3D is the problem that applies computer vision techniques to obtain the form of the human body from a given single image or a sequence of images. This problem has recently received considerable attention from the scientific community. The main reasons for this trend are the growing new areas of applications which are driven by the latest technological advances. Although recent approaches and the advent of deep learning in computer vision have presented remarkable results, 3D pose estimation remains a commonly unsolved issue.

Reconstruction from monocular RGB image is excessively more challenging since strong self-occlusions and the inherent depth ambiguity cause a very ill-posed reconstruction problem. Deep learning techniques contribute to the performance of current state-of-art solutions to more acceptable levels. Despite all issues and the inherent difficulty of the problem, upcoming techniques and observations can help to present better results in benchmarks.

1.1. Key Contributions

In this thesis, we utilize a multi-view camera environment to obtain accurate 3D human poses. First, we simultaneously recover 3D poses from each image coming from

different views. For this purpose, we make use of the state-of-the-art HMR method [7] as a single-view 3D human pose estimator. The HMR method is chosen because it can recover human pose from in-the-wild images and has very competitive performance in standard benchmarks like the Human3.6M protocol I-II [1]. After collecting 3D poses from each image frame, we combine 3D pose information from multiple cameras using the Procrustes Analysis.

The contributions of this thesis can be listed as follows:

- *Procrustes Analysis based 3D pose reconstruction:* We propose a statistical shape alignment based reconstruction method for human pose estimation. In this method, predictions coming from a multi-view environment are aligned to achieve more accuracy in a pose.
- *Camera Selection:* We propose a naive approach to select camera locations to achieve better accuracy than random selection. We also explain the effect of purposedly placing cameras on the accuracy of pose estimation.
- *Dynamic Camera Selection:* We figure out how much improvement can be obtained in the scenario that camera selection is made frame by frame. We also propose body-orientation based dynamic camera selection. However, this method does not contribute to overall reconstruction accuracy as we expect.

Overall, we propose *Procrustes-Analysis based 3D human pose estimation method*, which combines the poses obtained from a multi-view setup to achieve more accuracy. Procrustes Analysis uses the multi-view information, which is more robust to the challenges of 3D pose estimation.

1.2. Thesis Outline

First, Chapter 2 presents a literature survey of the existing methods, models, and datasets to provide a summary of current and previous studies in the context of our research.

In Chapter 3, we broadly explain the datasets used in this thesis. We also introduce our preprocessing algorithm to handle different datasets.

Then, we explain our key contributions and their performance in the experiments in Chapter 4 and Chapter 5, respectively. Specifically, in Chapter 4, we propose our technique that reconstructs 3D human pose from multi-view images.

In Chapter 5, experimental scenarios, performance analysis of our method are presented. We also compare the performance of our approach with current and previous studies in the literature.

Lastly, in Chapter 6, we resume our contributions and review the results of the proposed method in the thesis. Finally, we conclude by proposing future research directions.

2. RELATED WORK

In this chapter, we summarize a literature survey of the existing methods, models, and datasets to provide a summary of current and previous studies in the context of our research. First, we briefly introduce the 2D/3D pose datasets. Then, the human body representation models are discussed by briefly describing their distinctive features. Lastly, we look at single-view and multi-view pose estimation approaches in the literature.

2.1. Human Body Pose Datasets

2.1.1. Human3.6M [1]

Human3.6M is a popular motion capture dataset and has well-defined 3D pose benchmarks. There are 3.6 million poses and corresponding images. Six male and five female actors perform 17 different scenarios (walking, discussion, eating . . .) in a multi-view camera environment. There are four calibrated high-resolution video cameras in the setup.

As we see in Figure 2.1, the cloths on actors are regular, opposing to special motion capture suits, to be more realistic. Ionescu *et al.* [1] utilize seven subjects (three females and four males) as the training and validation set, and four subjects (two males and two females) for the testing set.

3D pose annotations are skeleton-model based. The location 19 keypoints of the human body are annotated in 3D space. Annotations are measured in meter.

2.1.2. MPI_INF_3DHP [2]

Mehta *et al.* [2] propose this dataset captured in multi-view environment. The authors make use of a commercial markerless motion capture tool [8] to obtain ground

Figure 2.1. Sample images from the Human3.6M dataset, expressing the variation in subjects, poses and viewing angles [1].

truths. The background is chosen as green to augmentation and segmentation issues. There are eight actors (four males and four females). The actors perform eight activity sets involving complex poses. There are 14 high definition cameras from a wide range of viewpoints and different elevations. As the authors report, there are more than 1.3 million frames captured with 3D skeleton-based annotations. There exist training and test sets separately.

Figure 2.2. Sample images from test set of the MPI_INF_3DHP dataset.

2.1.3. CMU Panoptic [3]

The CMU Panoptic dataset is one of the richest multi-view cameras environment in literature for human pose problems. There are 31 HD cameras and 480 VGA cameras in the studio. The CMU Panoptic is publicly available with rich annotations. There are 3D annotations for body, hand, and face. The dataset consists of video sets involving a single actor and multiple actors. In this thesis, we only use the single actor video sets with 3D body landmark annotations. 3D annotations are in a skeleton-based form. There are also Kinect-based annotations, but in this thesis, we do not make use of them. 'range_of_motion' subset is suitable for 3D human pose estimation since there is a single actor in the environment with 3D body annotations. There is no given 2D annotation in the dataset, but 2D information is obtainable in the existence of 3D annotations and camera calibration parameters.

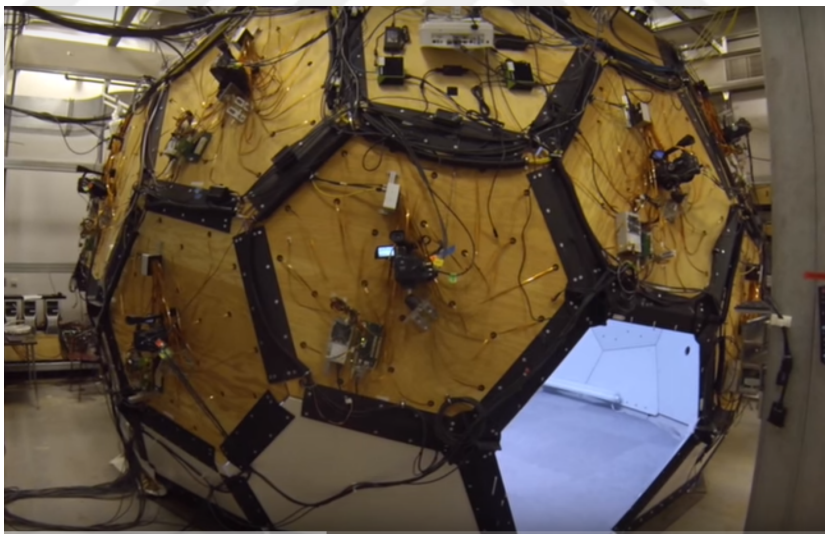

Figure 2.3. The CMU Panoptic motion capturing environment [4].

2.1.4. Datasets with 2D annotations

The MPII Human Pose dataset [9] has 25K images containing 40K people with 2D body joint annotations. There is no 3D information about human keypoints. The images are gathered from a large number of daily activities. Each image is obtained from a Youtube video.

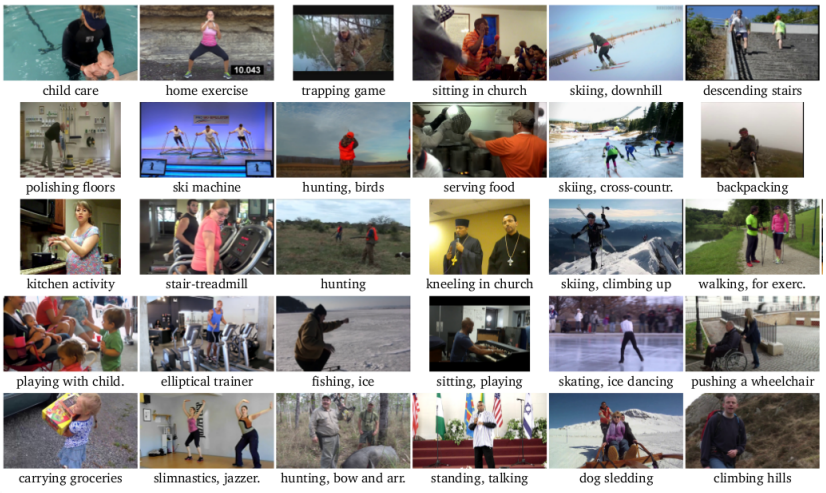

Figure 2.4. Sample images of some daily human activities in MPII dataset.

The COCO dataset [10] is large-scale in terms of annotations and images. There are more than 330K images and 250K people with 2D joint annotations.

Figure 2.5. Sample images with 2D annotations on the COCO dataset.

2.2. Human Body Representation Model

2.2.1. Skeleton Models

A pose is represented by the location of articulation points (neck, knees, elbows, etc.) in 3D coordinate space. This representation model is generally sufficient for action recognition problems. Likewise, the model is the simplest way to define human body configuration. There is no standard in articulation point count in literature. The point count can commonly change from dataset to dataset. A sample of the skeleton model is seen in Figure 2.6.

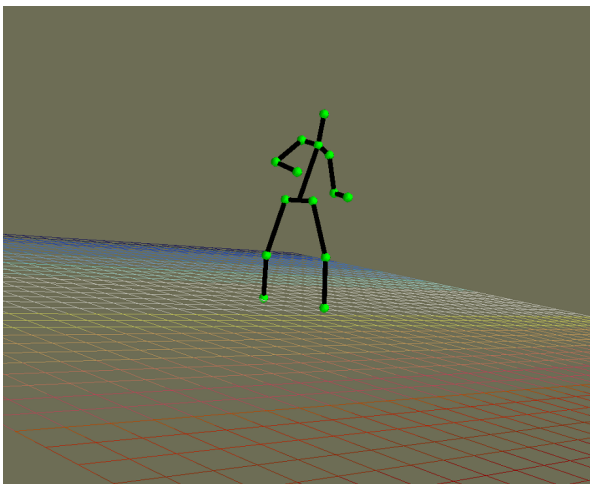

Figure 2.6. The skeleton model representation of human body pose.

2.2.2. 3D Shape Models

As mentioned before, recovering 3D human pose is a hard problem. Since there are too many independent variables in a human pose, it is always attractive to convert the problem into a well-defined parametric shape body model. In early literature, there are many 3D shape models to tackle the problem, as seen in Figure 2.7.

In recent literature, current models, such as SMPL [11], DMPLx2, are more representative to show body deformations and shape differences. SMPL is a learned model of human body shapes, trained a large number of 3D meshes of different people in varying poses. The model is vertex-based, and pose depended blend shapes are linear functions of pose rotation matrices.

2.2.3. Volumetric Models

In volumetric representation, the human body is modeled as the voxel grid in the 3D space. The 3D space is divided into fixed dimensions. Spaces belonging to the human body are marked as one encoded voxels. All other spaces are encoded as zero encoded voxels. Sample human body representation with volumetric model is given in Figure 2.8.

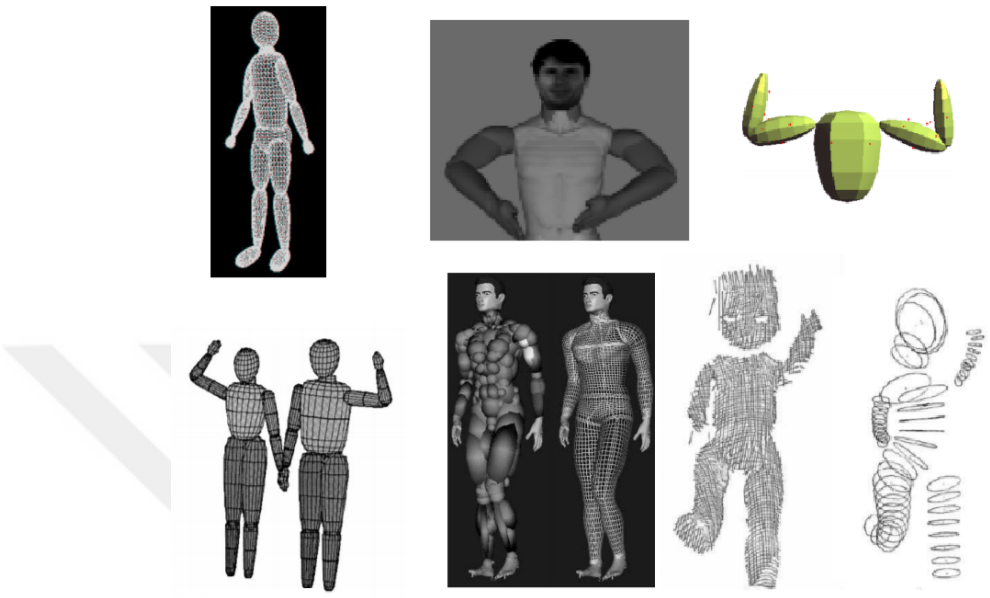

Figure 2.7. Early shape models.

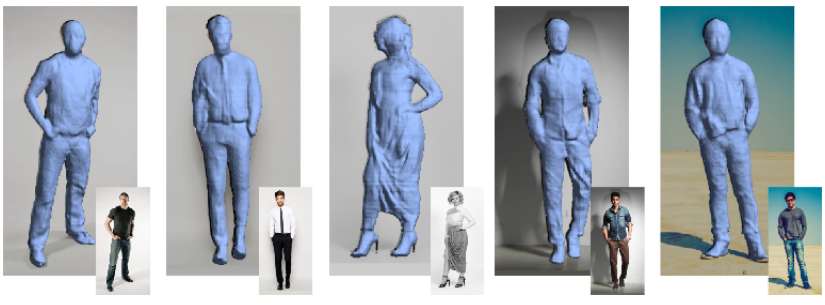

Figure 2.8. Volumetric representation of human body [5].

2.3. Methods

In the context of 3D pose estimation from a single image, techniques with a convolutional network trained end-to-end are generally dominates the performance benchmarks in recent years. Approaches using multi-view images are utilized for achieving better accuracy or for handling more challenging poses and environments.

2.3.1. Single-view Methods

In the literature, some techniques utilize a 3D shape model to constitute a 3D pose. Earlier methods used basic shapes; however, more recent methods use complex parametric shape models like SMPL. The deformations and variations in the human body are represented with parameters. During defining the pose estimation problem, this shape model is fitted in an optimization problem so that the projection of a model into the human in the image will correspond regarding a defined loss.

In Simplify [12], Bogo *et al.* use automatically detected 14 2D keypoint locations, then use an optimization method to minimize the difference between 2D keypoints and projection of the fitted SMPL model. Lassner *et al.* [13] upgrade Simplify by adding more fitting objective (91 2D keypoints used).

Tung *et al.* [14] propose a framework to directly recover the parameters of the 3D model using a deep neural network (DNN). The DNN is trained with the direct supervision of synthetic 3D data and weak-supervision of 2D landmarks.

Pavlakos *et al.* [15] combine the 2D landmark heatmaps and silhouette information to produce SMPL model parameters. A differentiable renderer is used to reproject the 3D shape model to 2D landmarks and silhouette. By maintaining differentiability, the authors can produce 3D mesh from estimated model parameters and optimize the surface by using 3D vertex-based loss. By keeping differentiability in the renderer, their model can be trained end-to-end.

Kanazawa *et al.* [7] propose Human Mesh Recovery (HMR), an end-to-end framework for constructing a full SMPL-based 3D mesh of a human body from a single image. HMR uses unpaired data: 2D keypoint annotation dataset and a separate dataset of 3D meshes to get around the lack of large-scale ground truth 3D annotation in the wild. Moreover, the authors propose a discriminator network to prevent the generator network that predicts the body parameters, from getting anthropometrically unpleasant results.

In NeuralBodyFitting, Omran *et al.* [16] utilize body segmentation parameters to estimate SMPL parameters. In DenseBody, Yao *et al.* [17] propose to utilize UV position map to predict SMPL model parameters. The authors directly predict SMPL model parameters without relying on any 2D representations.

In [18], Zhou *et al.* utilize 2D and 3D annotations with a weakly-supervised transfer learning technique. Their model is trained in an end-to-end manner and takes advantage of the relation between the 2D pose and 3D depth estimation. As a result, the 3D pose information gathered in a controlled environment is transferred to images in-the-wild.

Some researchers propose methods that estimate 3D pose directly from images. Tekin *et al.* [19] train an autoencoder to find out the latent representation of human pose, and then directly regress 3D human poses from 2D images.

Pavlakos *et al.* [20] utilize regression of 3D heatmaps instead of 3D coordinates. A two-step technique is employed: a ConvNet estimates 2D keypoint locations and an optimization process to recover 3D poses. The ConvNet predicts voxel likelihoods of each keypoint. In Integral Pose Regression, Sun *et al.* utilizes combined volumetric heatmaps with *soft-argmax* activation.

Some methods in the literature estimate a 3D pose in two-stages. First, 2D keypoints are detected with a 2D pose estimator. Then, to obtain a 3D pose, there are some methods utilized: regression, model-fitting, dictionary lookup from 3D pose pool,

etc. Two-stage approaches are more robust to environment shift, but they excessively rely on the accuracy of 2D joint estimation. Another substantial disadvantage of the two-stages approaches is not to make use of image information during 3D pose estimation. However, the latest methods in the literature are commonly end-to-end trained, directly makes use of image information to obtain a 3D pose.

Adversarial learning is powerful to distinguish samples of a source domain from samples of another domain. Yang *et al.* [21] make use of adversarial learning to rectify the 3D human pose structure from large scale mocap datasets to the images in-the-wild with only 2D pose annotations. The authors propose a discriminator network to differentiate the predicted poses and the ground truth poses. By the end of the training, discriminator enforces the generator to produce more reasonable poses even in the wild images.

Motion capture datasets are hard to produce since framework cost and inevitable nature of the indoor setup. These datasets can contribute to inference in the training of motion structure and dynamics, but not contribute well to the uncontrolled environment. In the literature, there are some works that are less depended on the direct supervision of mocap datasets. They need less 3D annotated data or only need 2D annotated data to train a 3D pose estimator model.

2.3.2. Multi-View Methods

Many approaches [22], [23], [24] uses the pictorial structures model [25] to recover 3D pose taken from multiple cameras. Pavlakos *et al.* [26] use a generic convolutional network for 2D pose estimation and obtain 2D pose heatmaps for each view. Then, by employing a 3D pictorial structures model, single view prediction heatmaps are combined to estimate 3D pose. However, as stated in [26], an application of the basic pictorial structures model in 3D has high computational cost because of the six degrees of freedom.

In [6], Rhodin *et al.* propose multi-view constraints as weak supervision. Their approach makes it possible to train a deep neural network to predict 3D pose for actions with little labeled data. The system is forced to predict the same pose in all views as in Figure 2.9. In their experiments, it has shown that a small dataset of images with corresponding 3D poses can be effectively complemented by a bunch of images obtained from multiple synchronized cameras, even if the relative positions of the cameras are not specified.

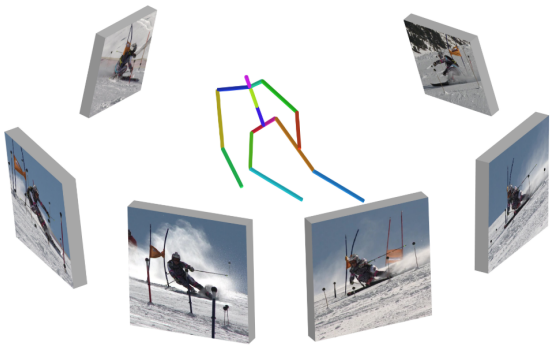

Figure 2.9. Multi-view constraints as weak supervision in [6].

In BodyNet [27], Varol *et al.* fits SMPL to ConvNet volumetric outputs as a post-process step. A synthetic dataset of rendered SMPL bodies is utilized for training a convolutional model for depth and body part segmentation. The authors also propose the multi-view re-projection loss to cope with the complexity in the human body articulation. When deep–net trained with re-projection losses, the performance increases both with single-view constraints. They show that the multi-view re-projection loss emphasizes on the body surface, which yields a better SMPL fit.

In EpipolarPose [28], Kocabas *et al.* present a self-supervised learning method, which does not need 3D annotations or camera extrinsic. In the training phase, the method estimates 2D poses from multi-view images, and obtain 3D pose and camera geometry by using epipolar geometry. Thereafter, the 3D pose estimator is trained with self-obtained 3D poses and camera parameters.

3. DATASETS

In this chapter, we mention the mocap datasets in detail. Then, we give statistics about datasets. Also, the details of preprocessing on the datasets are described.

3.1. Datasets Preprocessing

In general, there is no standardization in annotation techniques. So, it is necessary to preprocess the datasets that are planned to be used in a shared format before using. In this thesis, we process all the datasets in a standard format to feed them to the Deep Learning model and evaluate the performance.

In Figure 3.1, we plot the denotation of the 14 joints based skeleton-model used in this thesis. We transform the ground truths of skeletons for all datasets into 14 joint-based skeleton configuration.

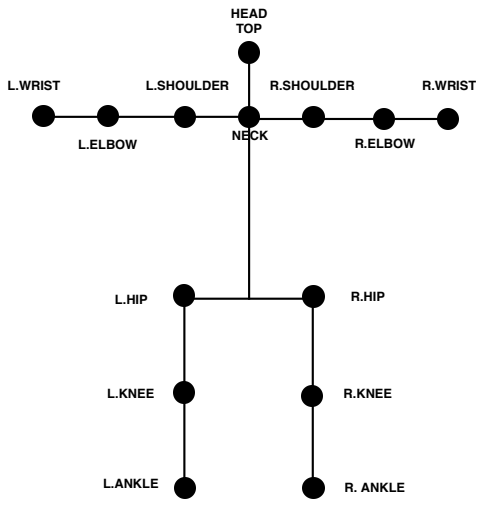

Figure 3.1. The denotation of keypoints in human body.

Table 3.1. Statistics about the CMU Panoptic 'Range of Motion' sequences.

Sequence Name	**Duration**	**# of Frames**	**# of Annot'ed Frames**	**# of Cams**
171204_pose1	17:30	31661	27561	31
171204_pose2	22:30	40689	37751	31
171204_pose3	5:00	9204	8920	30
171204_pose4	17:30	31671	31397	30
171204_pose5	15:00	27200	26902	30
171204_pose6	12:50	23301	22886	30
171026_pose1	13:20	24220	22466	31
171026_pose2	9:00	16366	14974	31
171026_pose3	4:20	7994	7181	31

3.1.1. CMU Panoptic

There are too many action sets in the CMU Panoptic. This dataset consists of video sets involving a single actor and multiple actors. We use only the single actor video sets with 3D body landmark annotations. Among all sequences with body annotations, we choose only 'Range of Motion' sequences.

In Table 3.1, duration, number of frames, number of annotated frames, and number of available HD cameras are given for corresponding motion sequences. In our experiments, we use each 30th frame of sequences.

In Figure 3.2, we plot the simultaneous frames of all HD cameras. The camera angles and elevations are highly diversified. As seen in Figure 3.2, *Camera21* is green. For some sequences, *Camera20* is not published. Therefore, we do not utilize *Camera20* and *Camera21* for all sequences for standardization concern.

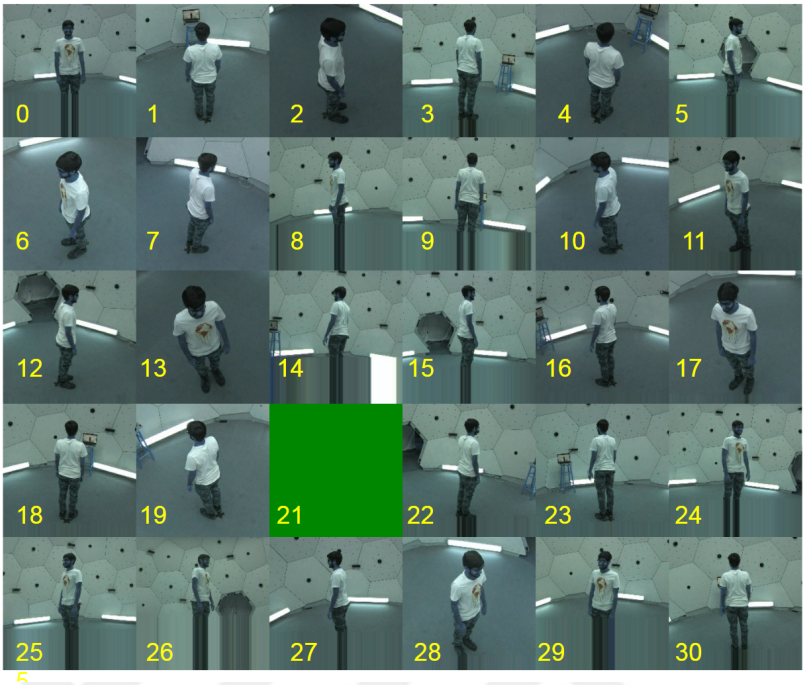

Figure 3.2. Frames from all HD cameras from different views in the CMU Panoptic.

Our preprocessing operations on the CMU Panoptic dataset are as follows:

(i) Convert HD videos to HD frames.

(ii) Check the existence of green or erroneous frames.

(iii) Get 3D body annotations, filter annotations with the incomplete skeleton.

(iv) Create 2D annotations by using camera parameters and 3D ground truths.

(v) Map the joint annotations of CMU Panoptic to our common joint denotation.

(vi) Filter the frames (like in Figure 3.3) with less than $MIN\_VISIBLE\_POINT$ (We set this parameter 8).

(vii) With proper frames, preprocessing is applied to the frames:

- Images are cropped so that skeleton stays in a tight bounding box with size 150 pixel.
- If necessary, images are padded to complete images to 224x224 frame size.
- 2D skeleton ground truths, camera parameters are updated regarding cropping and padding operations.

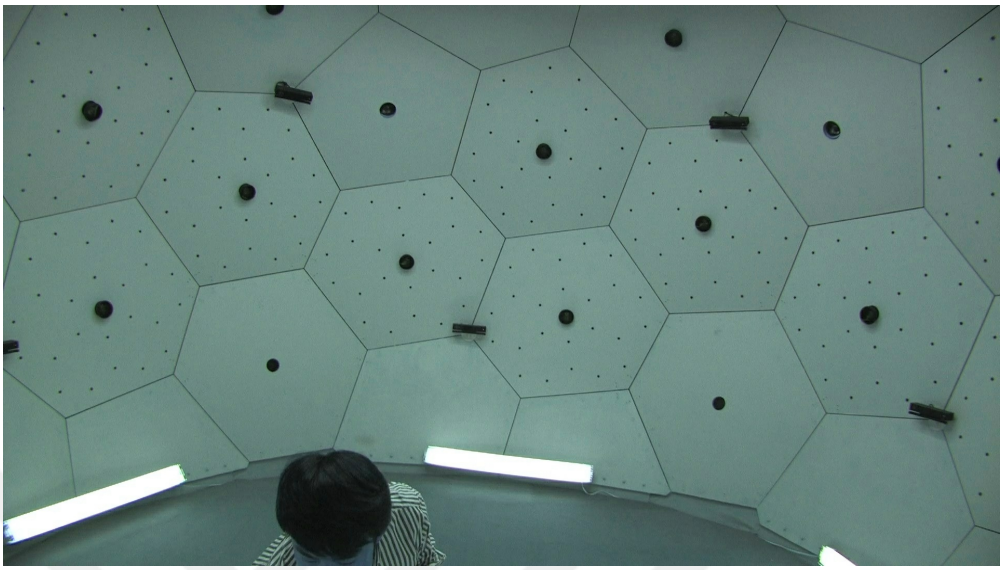

Figure 3.3. A sample skipped frame due to having fewer visible keypoints than the parameter of $MIN\_VISIBLE\_POINT$.

3.1.2. Human3.6M

There are 11 subjects: six male and five female actors. Four subjects are separated for testing; seven subjects are for training issues as referred to in Table 3.1.2. Each actor performs 15 different action sequences while four different cameras are recording from different viewpoints. In the Human3.6M setup, cameras are placed such that all keypoints of the actor are visible in the frame.

In the preprocessing, there is no need to check the number of visible keypoints. Different from the CMU Panoptic, there are also 2D annotations together with 3D annotations for each camera. The remaining processes are as follows:

(i) Convert HD videos to HD frames.

(ii) Images preprocessing are applied to the frames:

- Images are cropped so that skeleton stays in a tight bounding box with size 150 pixel.

Table 3.2. Subjects in Test/Train Set of the Human3.6M dataset.

Subject ID	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**
Set	Train	Test	Test	Test	Train	Train	Train	Train	Train	Test	Train
Gender	F	F	M	M	F	M	F	M	M	M	M

- If necessary, images are padded to complete images to 224x224 frame size.
- 2D skeleton ground truths, camera parameters are updated regarding cropping and padding operations.

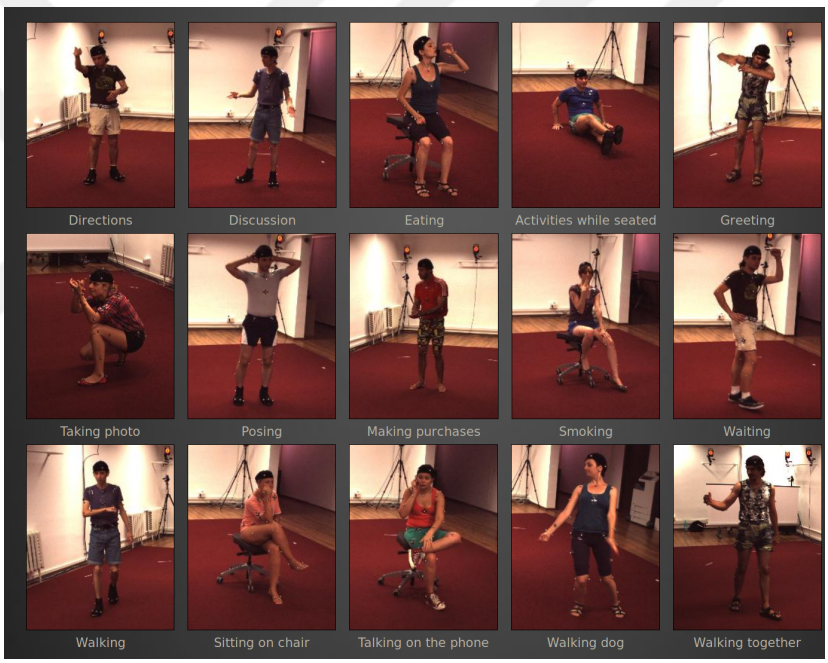

Figure 3.4. Action sets in the Human3.6M dataset.

3.1.3. MPI_INF_3DHP

There are eight different subjects in the training set. Each subject has two sequences involving 14 different camera views. Eight cameras are presented by default. Extra six cameras (three wall cameras and three ceiling cameras) are optionally obtainable. We do not process the frames of extra three ceiling cameras since we do not use in this thesis.

Our preprocessing operations on MPI_INF_3DHP dataset are as follows:

(i) Convert HD videos to HD frames (Resolution: 2000x2000).

(ii) Map the joint annotations of MPI_INF_3DHP to common joint labeling.

(iii) Filter the frames with less than $MIN\_VISIBLE\_POINT$ (We set this parameter 8).

(iv) With proper frames, preprocessing is applied to the frames:

- Images are cropped so that skeleton stays in a tight bounding box with size 150 pixel.
- If necessary, images are padded to complete images to 224x224 frame size.
- 2D skeleton ground truths, camera parameters are updated regarding cropping and padding operations.

Table 3.3. # of frames and cameras of MPI_INF_3DHP (subject, sequence) pair.

Subject ID	**Sequence ID**	**Frame Count**	**# of Camera**
S1	1	6416	14
	2	12430	14
S2	1	6502	14
	2	6081	14
S3	1	12488	14
	2	12283	14
S4	1	6171	14
	2	6675	14
S5	1	12820	14
	2	12312	14
S6	1	6188	14
	2	6145	14
S7	1	6239	14
	2	6320	14
S8	1	6468	14
	2	6054	14

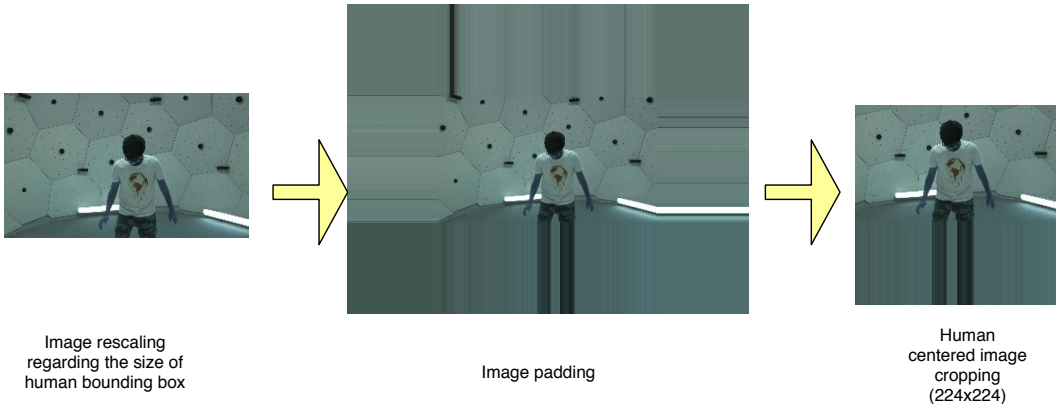

Figure 3.5. Image processing pipeline.

3.2. Bounding Box Preparation

The Deep Learning based pose estimators generally require the person in a well-centered tightly cropped images. We process the raw frames to prepare cropped frames with well-centered human body. In Figure 3.5, we simply plot our image processing pipeline. In detail, our processing method on frames are as follows in Figure 3.6:

Require IMAGE, Margin=150, gt2ds ;

Result $processed\_IMAGES, refined\_gt2ds$;

$min\_pt \Leftarrow min(gt2ds)$;

$max\_pt \Leftarrow max(gt2ds)$;

$person\_height \Leftarrow matrix\_norm(max\_pt - min\_pt)$;

$center \Leftarrow average(min\_pt, max\_pt)$;

$scale \Leftarrow 150/person\_height$;

$scaledIMAGE, scale\_factor \Leftarrow resize\_img(IMAGE, scale)$;

$height, width \Leftarrow shape(scaledIMAGE)$;

$scaled\_center \Leftarrow scale\_factor * center$;

$scaled\_gt2ds \Leftarrow scale\_factor * gt2ds$;

$start\_pt \Leftarrow max(center\_scaled - margin, 0)$;

$end\_pt \Leftarrow center\_scaled + margin$;

$end\_pt = [min(get\_width(end\_pt), width), min(get\_height(end\_pt), height)]$

$crop\_IMAGE \Leftarrow scaledIMAGE[start\_pt[1] : end\_pt[1], start\_pt[0] : end\_pt[0]]$

$refined\_gt2ds \Leftarrow scaled\_gt2ds - start\_pt$;

$processed\_IMAGES \Leftarrow PAD\_IMAGES\_224(scaledIMAGE)$

return $processed\_IMAGES$, $refined\_gt2ds$;

Figure 3.6. Bounding box operation for images by centering the human.

4. METHOD

4.1. Single View Pose Detector: Human Mesh Recovery

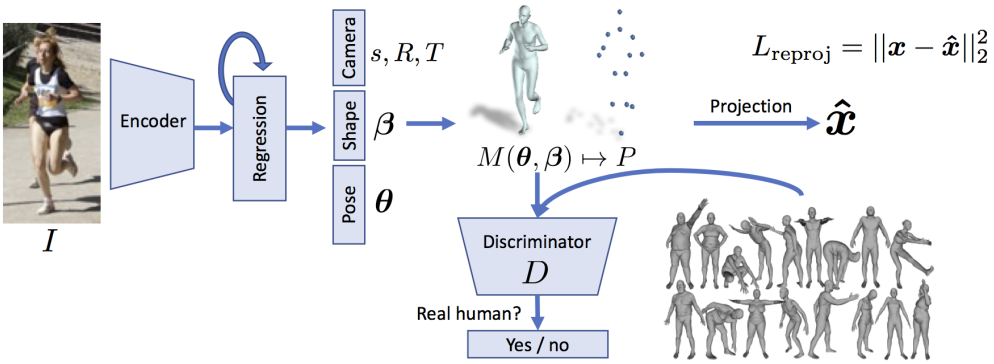

Figure 4.1. The overview of the HMR framework [7]

Kanazawa *et al.* [7] propose Human Mesh Recovery (HMR), an end-to-end system for estimating a full 3D mesh of a human body from a single image. The overall HMR framework is represented in Figure 4.1. Skinned Multi-Person Linear Model (SMPL) is used as a 3D mesh model in the HMR method. The HMR method minimizes the reprojection loss of joints, which allows the model to be trained using images that only have ground truth 2D body annotations. In the HMR method, a 3D pose model can be constructed with and without using any paired 2D-to-3D supervision. In our research, we utilize the 3D joint predictions, not 3D mesh.

4.2. Multiview Reconstruction

In this thesis, we propose a 3D human body reconstruction method that uses multiple cameras. We use a state-of-the-art single-view 3D pose estimator, namely the HMR method, which produces 14 3D body joint coordinates from a single image [7]. In a multi-view camera setup, we obtain estimates of body coordinates from each view. The integration of a set of joint positions to obtain the final reconstructed 3D model is carried out by using the Procrustes Analysis. Procrustes Analysis enables us to find

the best alignment of a set of ordered 3D points to a consensus shape. Consensus shape is defined as the average shape of the initial 3D estimations obtained from a single-view pose estimator. After the alignment of single-view 3D body coordinates to an average body model via the Procrustes method, we apply the component median method to the set of 3D joint coordinates to compute the final position of the combined 3D body joint. Component median method for a set 3D vectors finds the median of each (x, y, z) component. The use of the median approach is motivated by the fact that some of the individual 3D joint estimations produced by the single-view pose estimator may be unreliable. Thus, a statistically robust approach, e.g., use of the median method, is needed to compute the final average position of a 3D body joint from a set of initial estimates. The overall structure of the proposed approach is illustrated in Figure 4.2.

Obtaining the consensus 3D human shape from a set of single-view poses is one of the most critical phases in our proposed approach. We utilize the Procrustes Analysis, one of the primary statistical shape analysis method from the image analysis point of view, to reconstruct accurate "combined" 3D human pose. Procrustes Analysis uses the Procrustes distance measure and minimizes this distance to align two poses by finding the best translation, rotation, and scaling parameters. Procrustes distance is given as (4.1):

$$\min_{s,R,t} \left\{ \frac{1}{J} \sum_{i=1}^{J} \|y_i - (sRx_i + t)\|^2 \right\} \tag{4.1}$$

where s is a scaling factor, R is a rotation matrix and t is a translation vector, J is the number of joints, y_i is the target pose, x_i is the pose to be aligned.

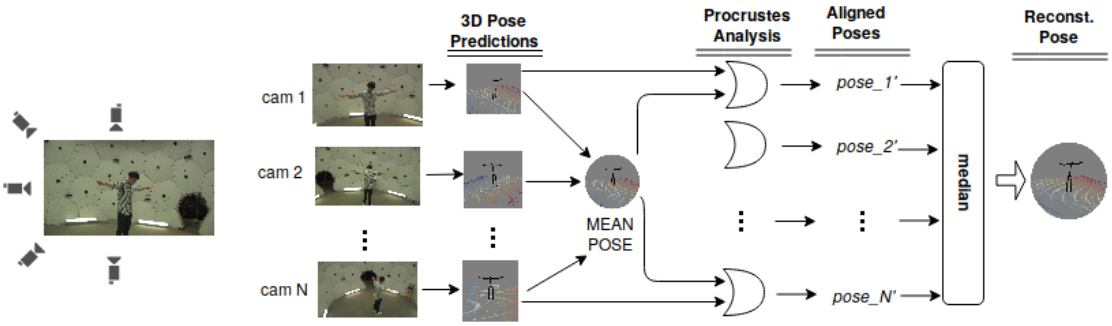

Figure 4.2. The cameras from different viewpoints, simultaneously generate images. The chosen 3D pose estimator, HMR predicts 3D pose of the human in the scene for each image. Then, the mean of predicted poses is calculated. All the poses are aligned with the mean pose by Procrustes Analysis. Finally, median filtering is applied to all aligned poses to reconstruct a "consensus" pose.

Input: Images from different cameras for time t

Output: 3D locations of body key points

n: usable camera count in the scene
select m cameras from all $m \leq n$
for all time t **do**
    fetch I_i from $cam_i, i \in CAM_m$
    for all I_i **do**
        calculate $pose_i$
    end for
    $avgPose_t = \frac{1}{m} \sum_{i=1}^{m} pose_i$
    for all $pose_i$ **do**
        calculate $procrustPose_i = Procrustes(pose_i, avgPose_t)$
    end for
    consensusPose = median($procrustPose_0$,.,$procrustPose_m$)
end for

Figure 4.3. Multi-view reconstruction algorithm

5. EXPERIMENTS AND RESULTS

In this chapter, the performance evaluation of our reconstruction method is explained. First, we describe the metrics. Then, the single view performance of each camera is analyzed on each dataset. Experimental scenarios are described, then, performance metrics are listed, and explained in detail. We examine the parameters of the reconstruction approach.

5.1. Evaluation metrics

We report the mean per joint position error (MPJPE) [1] which calculates the average Euclidean distance between ground truth and prediction for all keypoints.

$$MPJPE = \frac{1}{J}\sum_{j=1}^{J}\|pose_j - GT_j\|_2 \tag{5.1}$$

where $pose_j$ is predicted position of the joints; GT_j is correspondent ground truth of the joints.

We also use the rigid alignment of the prediction pose regarding the ground truth using Procrustes Analysis. This error metric is referred to as PA-MPJPE in literature. PA-MPJPE is better in evaluating the quality of the reconstructed 3D pose (skeleton) since global misalignments are removed. Both MPJPE and PA-MPJPE are measured in millimeters.

$$PA\text{-}MPJPE = \frac{1}{J}\sum_{j=1}^{J}\|alignedPose_j - GT_j\|_2 \tag{5.2}$$

where $alignedPose_j$ is Procrustes alignment of the predicted joints to the correspondent ground truth of the joints.

5.2. Experiments on the CMU Panoptic

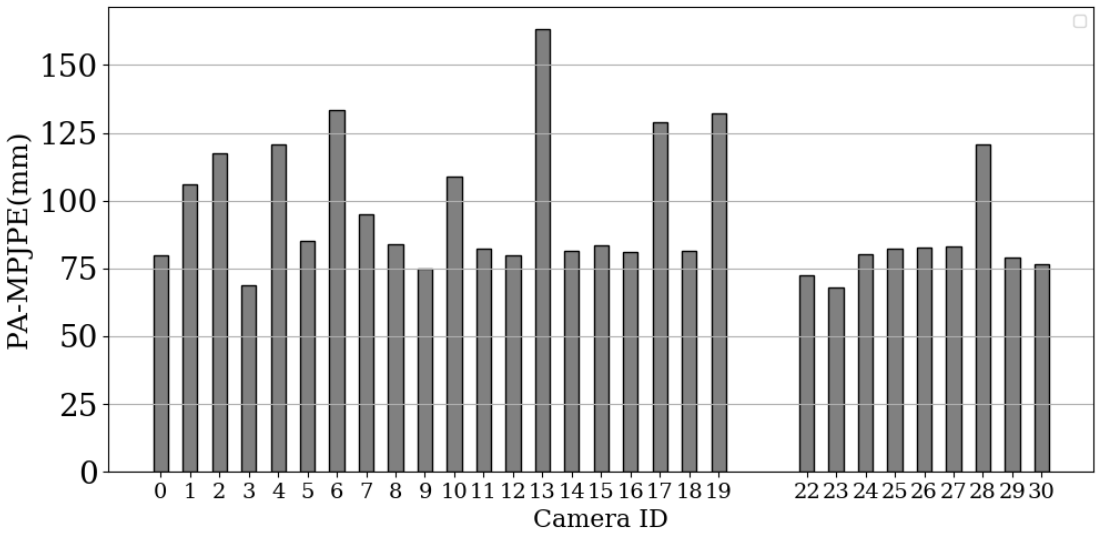

Figure 5.1. Reconstruction error for each camera on the CMU Panoptic.

In Figure 5.1, we present the average reconstruction error for each single camera, excluding Camera 20 and 21. There are many extreme camera angles in the CMU Panoptic Studio. We have observed that the performance of HMR is low on extreme camera angles since the pre-trained model of HMR is not trained on cameras with upper viewpoints. Hence, we do not utilize the cameras that are placed higher than 3 meters. Excluded upper cameras are given in Figure 5.2. In our experiments, we also do not use Camera 20 and Camera 21 since there are some technical issues (i.e., full green record or not publicly available) in the motion sequences: *171204_pose3*, *171204_pose4*, *171204_pose5*, *171204_pose6*. Our last significant result in Figure 5.1, single-camera performances are not excessively variated among the remaining 20 cameras, which also demonstrates that HMR can run all horizontal viewpoints.

In Figure 5.3, we plot the camera views of the best three cameras in the top row. Camera 3, 22, 23 are installed diagonally at the height of 1.5-2 meters. The worst performances are obtained from Camera 6, 13, 19. In the bottom row of Figure 5.3, the views of the worst cameras are shown. We observe that the elevations of the

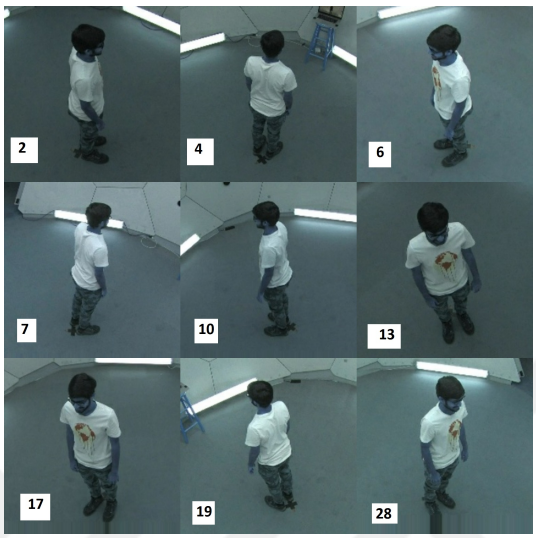

Figure 5.2. Excluded camera views of the CMU Panoptic in our experiments.

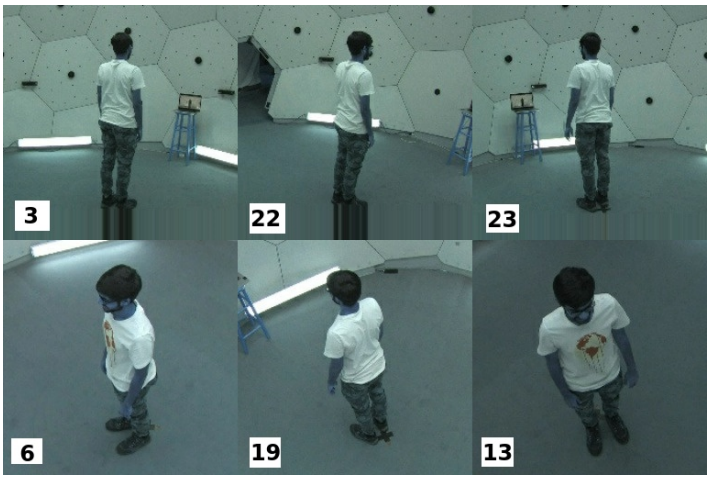

Figure 5.3. The top row: the best three cameras, the bottom row: the worst three cameras.

worst cameras are over 3 meters. Upper camera elevation leads to depth ambiguity and self-occlusion problems, which result in higher reconstruction errors.

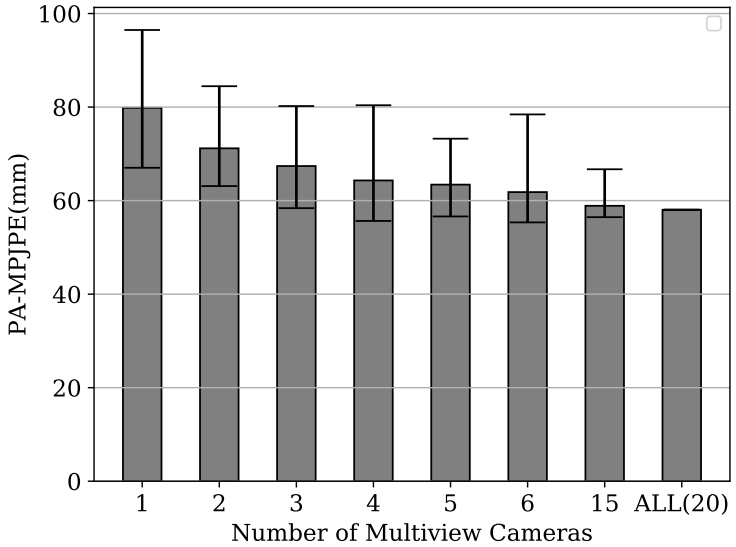

Figure 5.4. Reconstruction error for different multi-view camera setups. For each case, average, minimum, and maximum error rates are plotted, except for the *ALL(20)* case.

We utilize the remaining 20 HD cameras since upper cameras and Camera 20/21 are excluded from our experiments. In Figure 5.4, we show the mean, minimum, and maximum values of the reconstruction error (PA-MPJPE) in the case of a given number of randomly chosen cameras. As seen in Figure 5.4, increasing the number of cameras in the multi-view configuration improves the performance on the average. These results are calculated by exhaustively searching camera combinations for a given camera count. However, increasing the number of cameras more than four in a multi-view setup does not improve the best accuracy.

We also inspect the reconstruction filtering methods for *4CAMs* (4 Cameras) and *ALL(20)*. In Figure 5.5, we compare the performance of component-median, vector median, and average. For *ALL(20)*, there are only marginal differences in filtering methods.

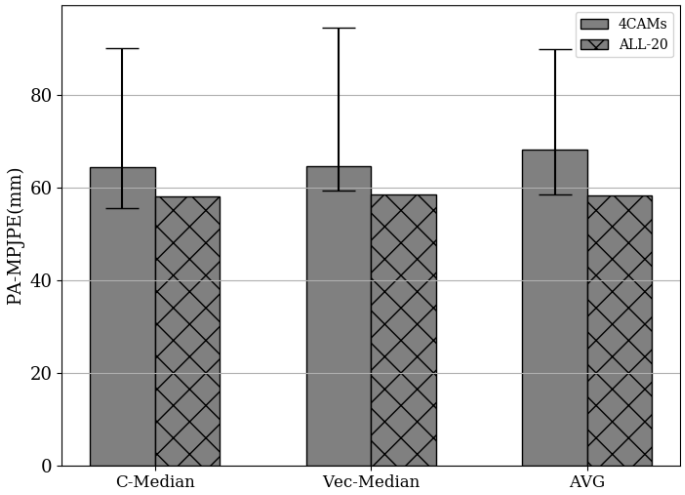

Figure 5.5. Reconstruction error with different filtering methods. For *4CAMs*, average, minimum, and maximum error rates are plotted, except for *ALL(20)*.

However, for the fewer number of camera configurations, there is no significant difference between the two median-based filtering methods. However, AVG (taking average of each component) is more exposed to the distorting effect of outliers than component-median and vector median. When we compare the two median-based filtering methods, component-median filtering is slightly better in the comparison of the average and the best performance.

In Figure 5.6, we analyze the performance of filtering methods on joints for the scenario that all cameras are utilized. The primary aim of this experiment is to figure out which method is better in especially in outer joints. However, there is no significant dominance of any method on outer joints.

5.2.1. Camera Selection

5.2.1.1. 4-CAMERA. In multi-view setups, like the CMU Panoptic, redundant information is high. If we can choose the configuration of cameras efficiently, the performance contribution of using more cameras may stay marginal. Using an extra camera is important because adding additional cameras results in increased setup and main-

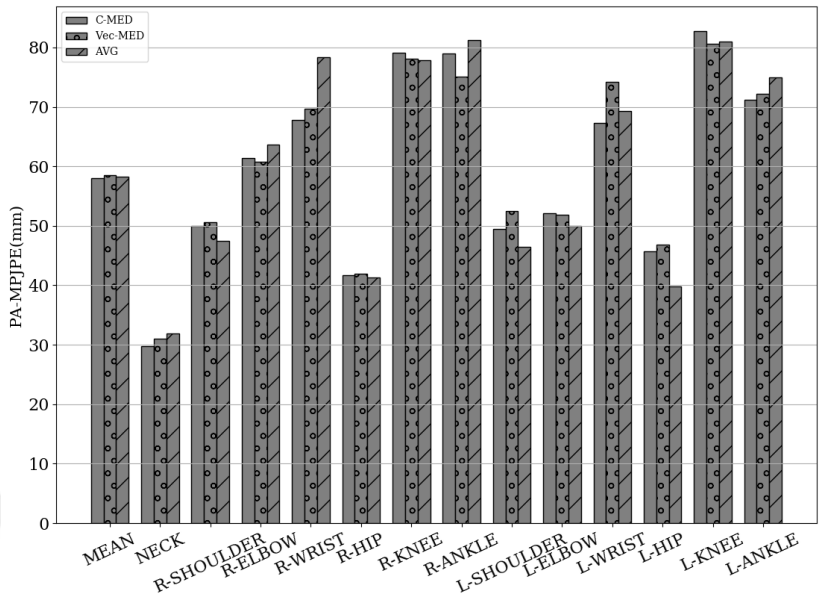

Figure 5.6. The joint-based effect of filtering methods on reconstruction error on the CMU Panoptic. (Using *ALL(20)*)

tenance costs. We analyze two different multi-view camera configurations: *Configuration1* chooses four cameras from each diagonal; *Configuration2* chooses four cameras from left, right, front, and rear sets. Proposed configurations are as seen in Figure 5.8. Regarding the results in Figure 5.8, *Configuration1* presents 3.0% better performance on average than *Configuration2*. A diagonally placed camera system produces a lower reconstruction error than *Configuration2*. The *BEST4* camera combination involves the four cameras that produce the least reconstruction error for test video. Therefore, *BEST4* depends on the test sequences. If we can choose the configuration of *BEST4*, the error will be 55.65mm. The performance gap between *BEST4* and *Configuration1* is only 4.1mm of reconstruction error. Moreover, *Configuration1* can produce only 1.7mm of error close to the reconstruction of *ALL(20)*.

In Figure 5.7, we observe that *4CAMs* with diagonal configuration show close performance to *BEST4* (impractical to know) and *ALL(20)* (costly) with a marginal loss. So, in Figure 5.9, we investigate the joint base reconstruction error of diagonally selected *4CAMs*. There is no joint-based performance improvement or significant degradation regarding *BEST4* and *ALL(20)*. All joints show close performance with a marginal difference.

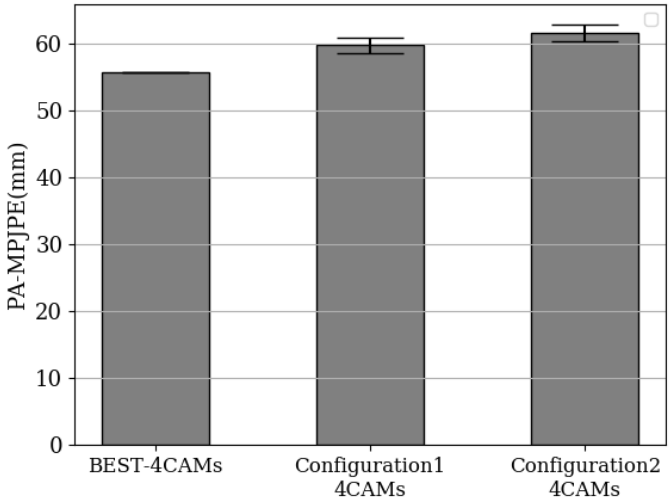

Figure 5.7. Reconstruction error (PA-MPJPE) for different *4CAMs* configurations. For *Configuration1* and *Configuration2*, variances of the errors are illustrated.

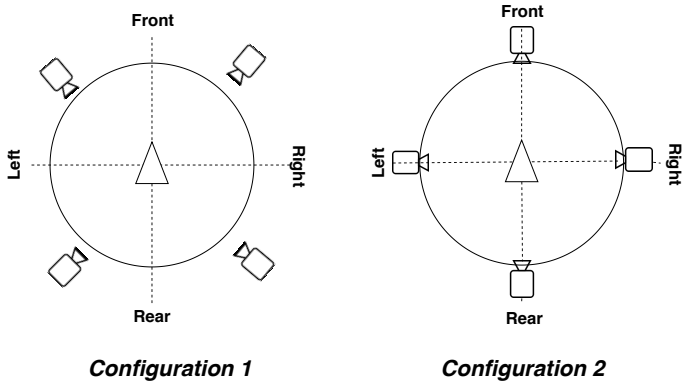

Figure 5.8. Multi-view camera (*4CAMs*) configurations: *Configuration1*, *Configuration2*.

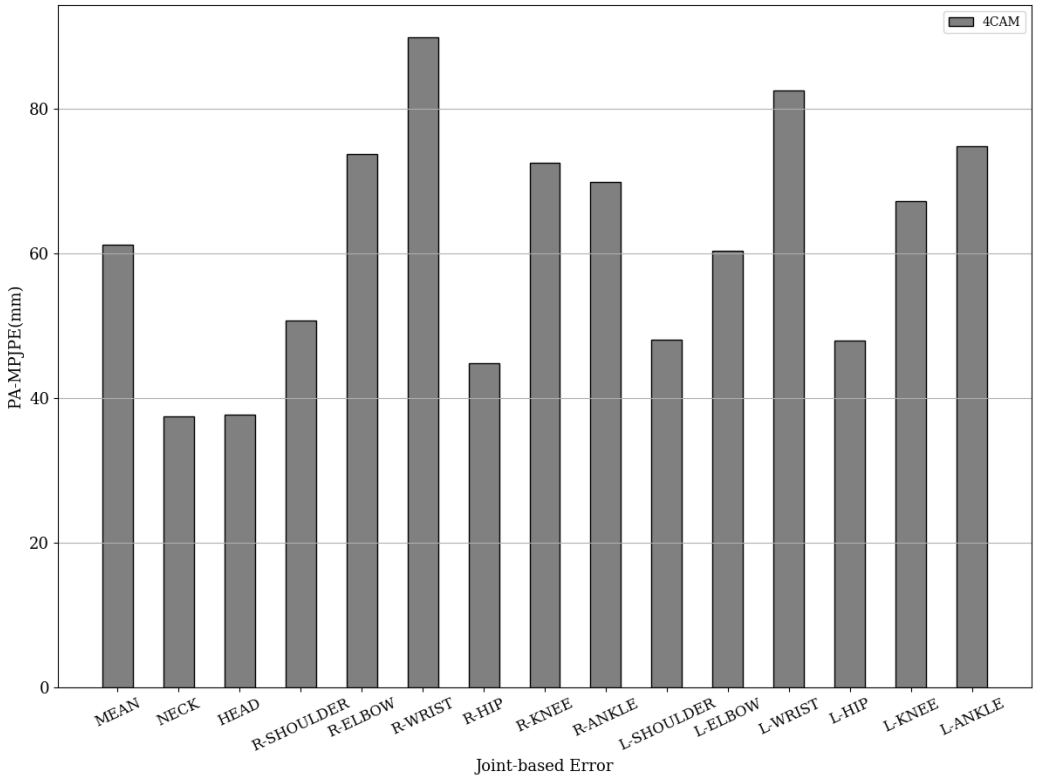

Figure 5.9. Joint-based reconstruction error analysis with diagonally selected *4CAMs*.

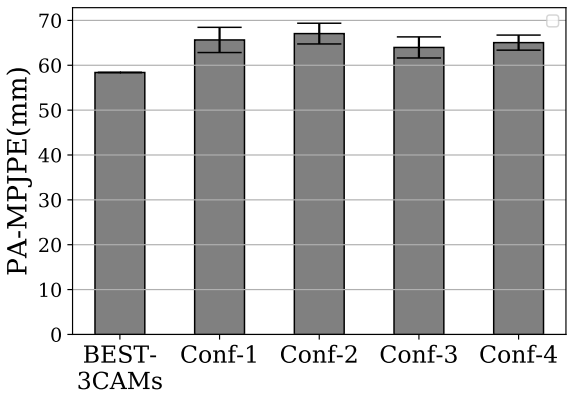

Figure 5.10. Reconstruction error (PA-MPJPE) for proposed *3CAMs* configurations. For *Configuration1*, *Configuration2*, *Configuration3*, and *Configuration4*, variances of the errors are illustrated.

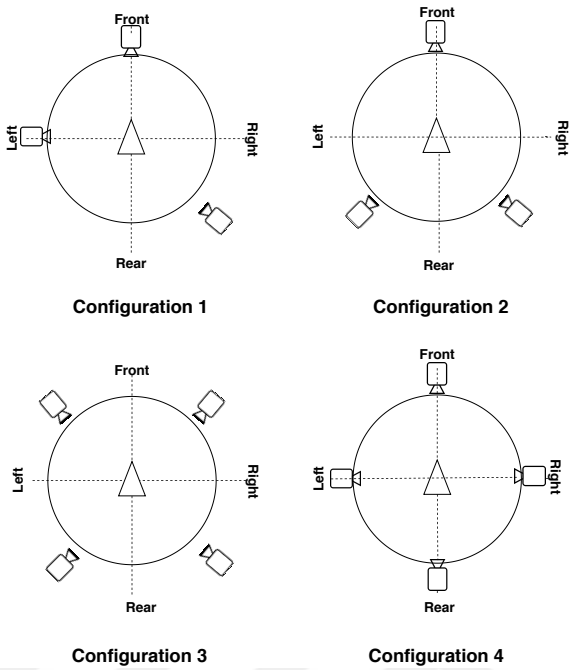

Figure 5.11. Multi-view camera (3CAMs) configurations: *Configuration1*, *Configuration2*, *Configuration3*, *Configuration4*

5.2.1.2. 3-CAMERA. In Figure 5.10, we analyze four different multi-view camera configurations: *Configuration1* chooses three cameras from front, left or right, and cross diagonal regarding selected side camera; *Configuration2* chooses two rear-diagonal cameras and front camera; *Configuration3* chooses three of four possible diagonal cameras; *Configuration4* chooses three of four cameras from left, right, front and rear sets. Proposed configurations are as seen in Figure 5.11. The performances of proposed configurations for three cameras are not close to the performance of $BEST3$. However, *Configuration3*, diagonally placed three cameras, is the most acceptable configuration among all four configurations. The configuration of *BEST3* can yield 58.39mm of error. *Configuration3* produces 5.58mm of error more than *BEST3*.

5.2.1.3. 2-CAMERA. In Figure 5.12, we analyze six different multi-view two-camera configurations: *Configuration1* chooses two cameras from front and side cameras (left/right); *Configuration2* chooses two mutually opposite diagonal cameras; *Configuration3* chooses two of rear diagonal cameras; *Configuration4* chooses two of front diagonal cameras; *Configuration5* chooses two front and rear cameras; *Configuration6* chooses side (left/right) cameras. The camera positioning in proposed configurations are as seen in

Figure 5.13. Our proposed poses are not close enough *BEST2*. But, *Configuration3* is the most acceptable configuration among all six configurations. If we can choose the configuration of *BEST2*, the error will be 63.11mm. On average, *Configuration3* can produce 2.13mm close to *BEST2*.

Configuration3 is highly acceptable to be chosen as a camera configuration. However, even though *Configuration3* and *Configuration4* are similar, *Configuration3* yields lower error since the actors in the studio perform close to front cameras and their lower bodies cannot be visible in the frames. In other words, *Configuration4* is more exposed to partial visibility of the human body. Unseen lower body results in high reconstruction errors in lower joints.

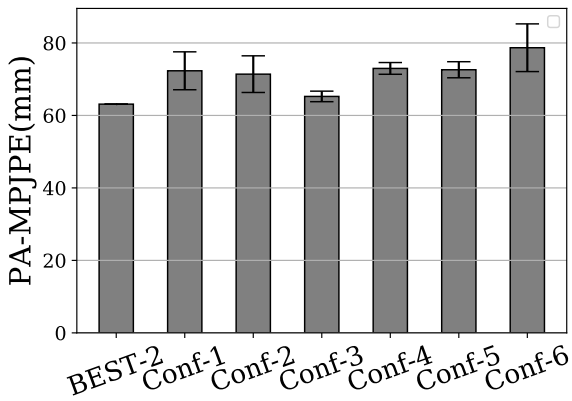

Figure 5.12. Pose estimation errors (PA-MPJPE) for different camera configurations. For *Configuration1*, *Configuration2*, *Configuration3*, *Configuration4*, *Configuration5*, and *Configuration6* variances of the errors are illustrated.

5.2.2. Dynamic Camera Selection

Up to now, camera configurations in our experiments are static. We utilize the same cameras until the end of the sequence. Dynamically selecting active cameras means possible changes in camera configuration frame by frame. In Figure 5.14, we analyze the performance contribution of dynamic camera selection frame by frame. We plot the reconstruction error for one, and up to five cameras, then we also plot using *ALL(20)*, which results in the same configuration for each frame since there is only one combination when all cameras are used.

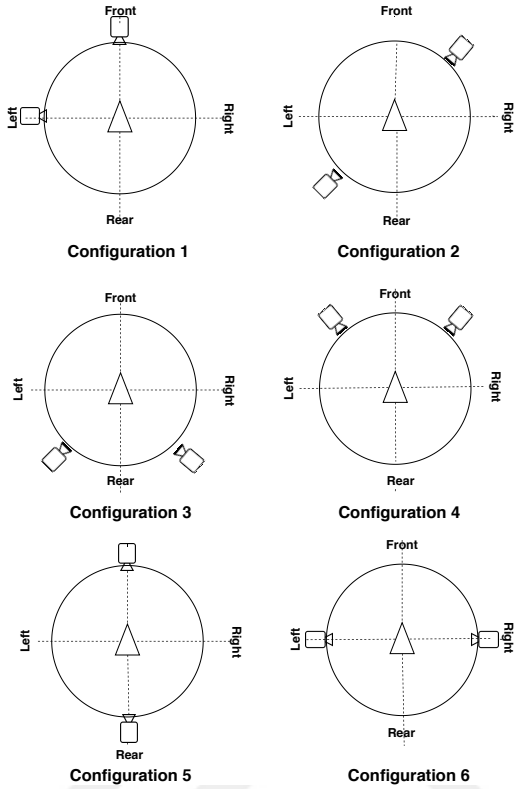

Figure 5.13. Multi-view camera (2CAMs) configurations: *Configuration1*, *Configuration2*, *Configuration3*, *Configuration4*, *Configuration5*, *Configuration6*.

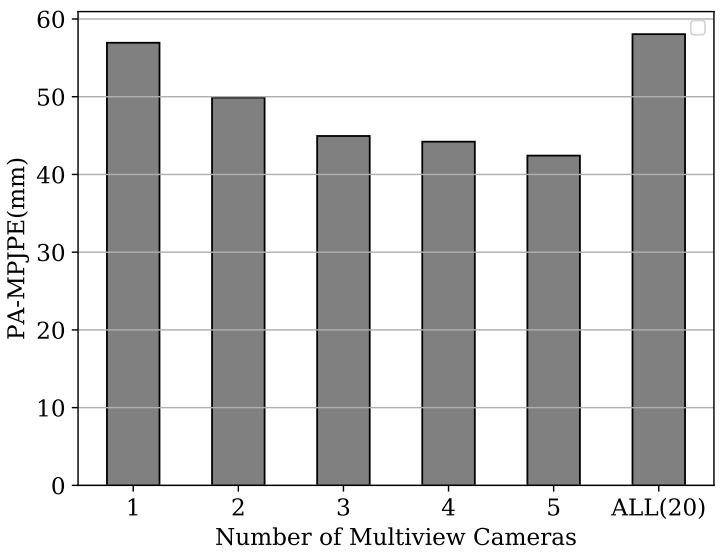

Figure 5.14. The reconstruction performance of dynamic camera selection with different number of multi-view cameras.

The motivation of dynamic camera selection is to re-configure cameras regarding the actor in the scene for achieving more accurate poses. As in Figure 5.14, dynamically selected *4CAMs* can yield 44.21mm of reconstruction error. Another significant result is that dynamically selected single-camera performance results in 56.94mm while *ALL(20)* can yield 58.05mm. However, these results are obtained by utilizing ground-truth information. It is not trivial to estimate the best n camera configuration that provides the lowest error for a given frame. Our experiments with dynamic selection aim to present the limits of our approach. It provides an idea of how further we can improve our results.

5.2.2.1. Body-orientation based Dynamic Camera Selection. In this section, we propose a naive approach for dynamic camera selection. As seen in Figure 5.15, we calculate the angle between shoulders and the line between neck joint and camera position from the bird-eye view. Body orientation is obtained by getting a perpendicular vector to the line of shoulders. We retrieve the position of neck and shoulders from ground-truth since our motivation is first to observe the contribution of the proposed approach.

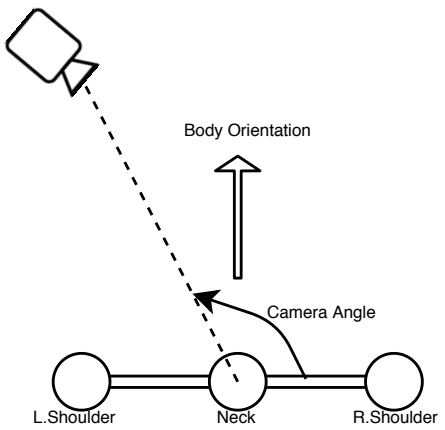

Figure 5.15. Calculation of body orientation and camera angle.

In the first scenario, we measure the performance of dynamic selection for the single-camera case. As in Figure 5.16, we propose four different camera positions regarding body orientation: *Configuration1* chooses a camera with orientation that is parallel with body orientation; *Configuration2* chooses a camera with orientation that creates intersects diagonally; *Configuration3* chooses a camera with orientation

that is perpendicular to body orientation; *Configuration4* chooses a rear camera with orientation that is parallel with body orientation.

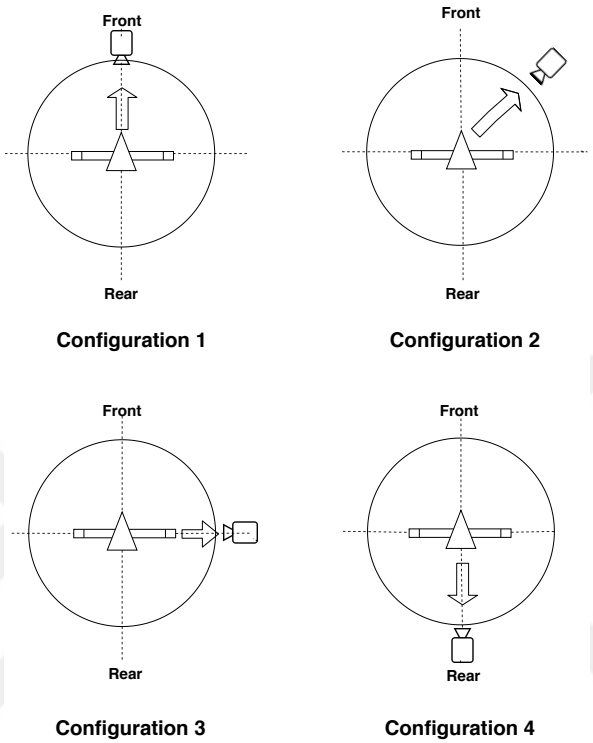

Figure 5.16. Different configurations for dynamic single camera.

In Figure 5.17, to analyze the performance of body orientation based dynamic camera selection, we compare the results of the experiments. *dynamic-1CAM* seems highly successful, but it is a symbolic lower-bound value to observe the other results. One of the significant results is that *Configuration2*, diagonally selecting a front camera, is the most successful among our proposed configurations. Another significant result is that *Configuration2* has almost same performance with *best-static-1CAM*. It is not trivial to know the best camera without any prior information about the motion sequence. *dynamic-1CAM* has significantly better performance than *Configuration2*, which is even the best configuration we propose. However, *Configuration2* seems to achieve similar performance to static best-camera. However, only using the supervision of body orientation does not yield similar performance to *dynamic-1CAM*, which is the ideal case, since the best camera configuration does not only depend on body orientation. Other factors are effecting the superiority of the camera, among all cameras. For example, the invisibility of exterior joints profoundly degrades the overall performance.

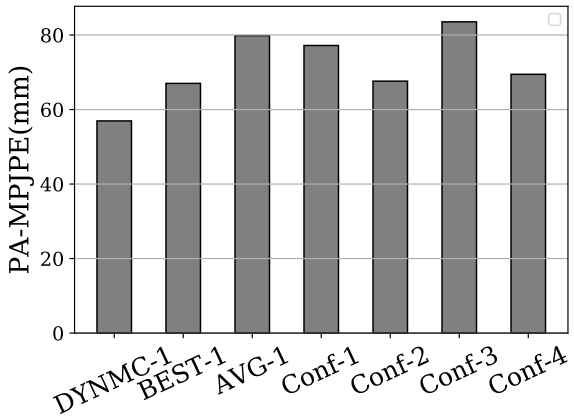

Figure 5.17. Performance analysis of proposed configuration for dynamic single camera selection.

In the case of four cameras utilized, we evaluate the same two configurations as used for static camera selection. We use dynamically these two configurations plotted in Figure 5.8: *Configuration1* chooses frame by frame four cameras from each diagonal regarding body orientation; *Configuration2* chooses frame by frame four cameras from left, right, front and rear regarding body orientation. While *Configuration1* yields 62.21mm reconstruction error, *Configuration2* yields 67.14mm of error. *Configuration1* achieves better reconstruction, but body-orientation based dynamic camera selection with four cameras does not improve reconstruction accuracy more than static diagonally placed four cameras.

5.2.3. Action Complexity

In Figure 5.19, we study the joint-based accuracy of our reconstruction method on actions from different complexity. As plotted in Figure 5.20, in the complex sequence, the actor performs a dance with extreme poses. Therefore, there are more articulation and self-occlusion in the poses in the complex action. Sample frames from simple action are plotted in Figure 5.21. In the simple action, the actor does not perform significant changes in body-orientation. As seen in Figure 5.19, exterior joints, like ankles and wrists, are exposed to more reconstruction errors than the inner joints, like shoulders.

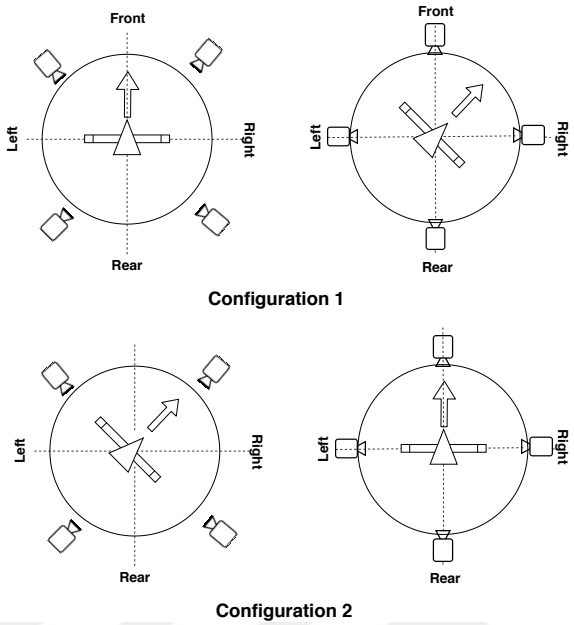

Figure 5.18. Regarding body orientation dynamically choose diagonal cameras (*Configuration1*), and dynamically choose perpendicular cameras (*Configuration2*).

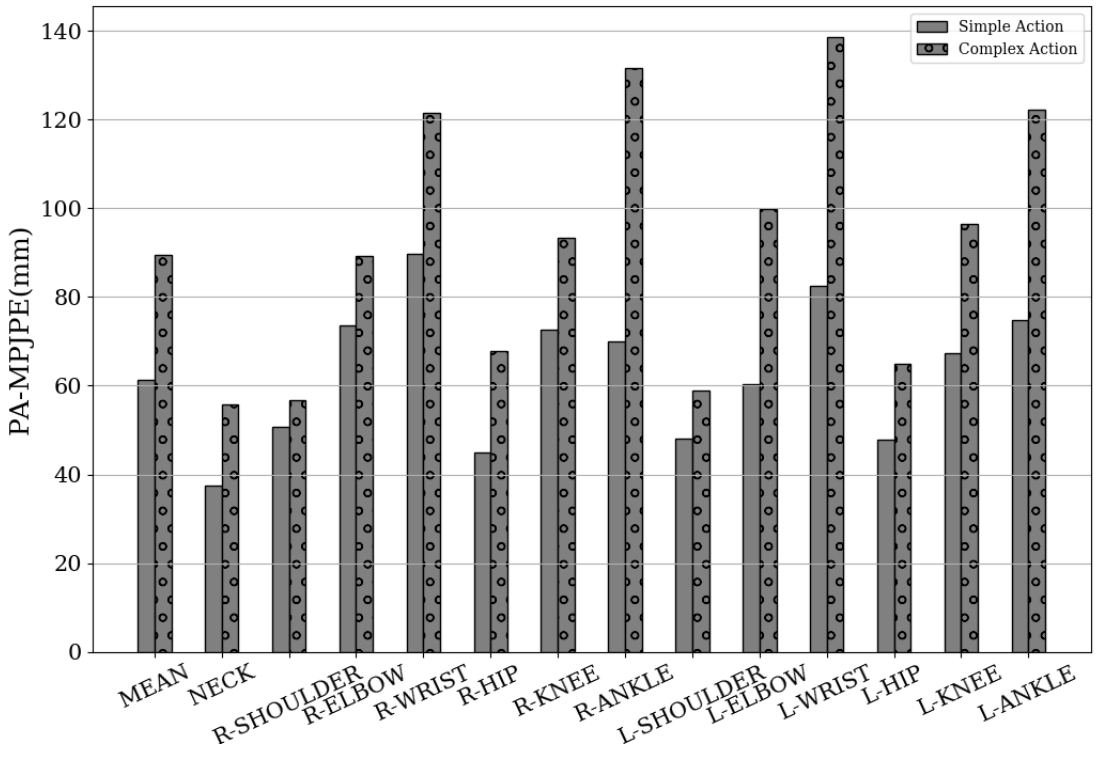

Figure 5.19. Joint-based error analysis on actions from different level of complexities (Simple Action vs Complex Action).

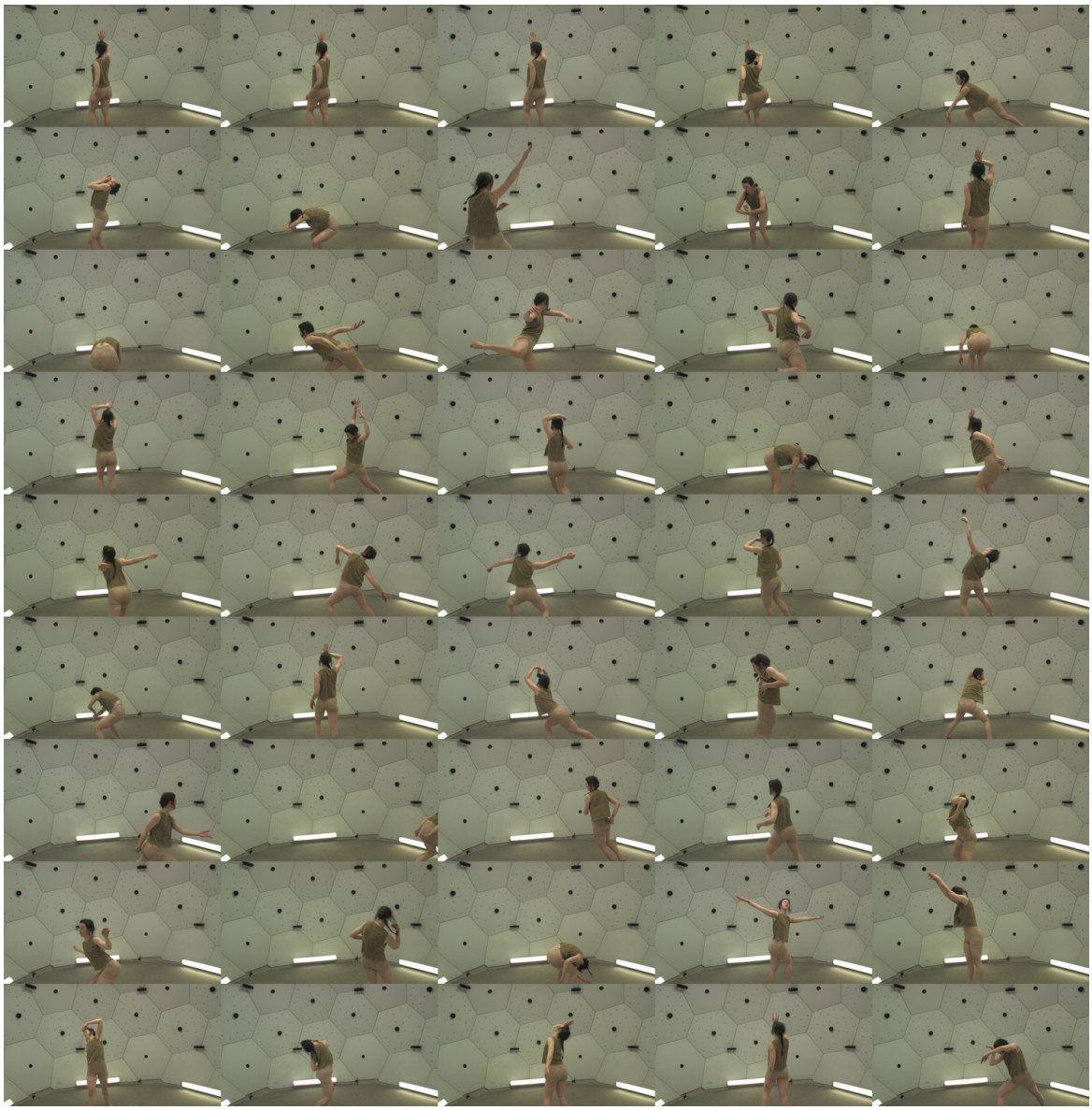

Figure 5.20. Frames from *170307_dance5* sequence on the CMU Panoptic (Complex Action).

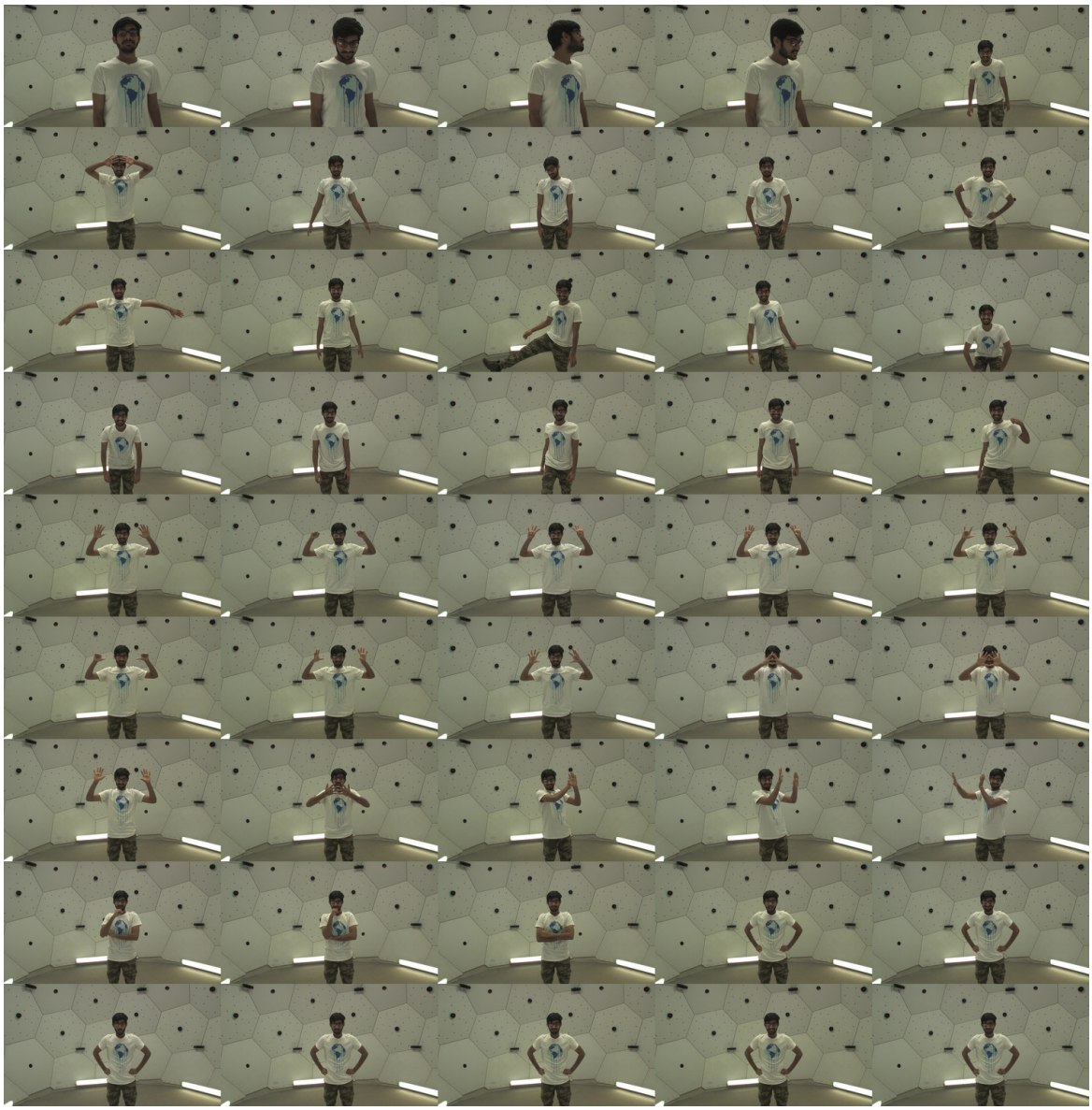

Figure 5.21. Frames from *171204_pose3* sequence on the CMU Panoptic (Simple Action).

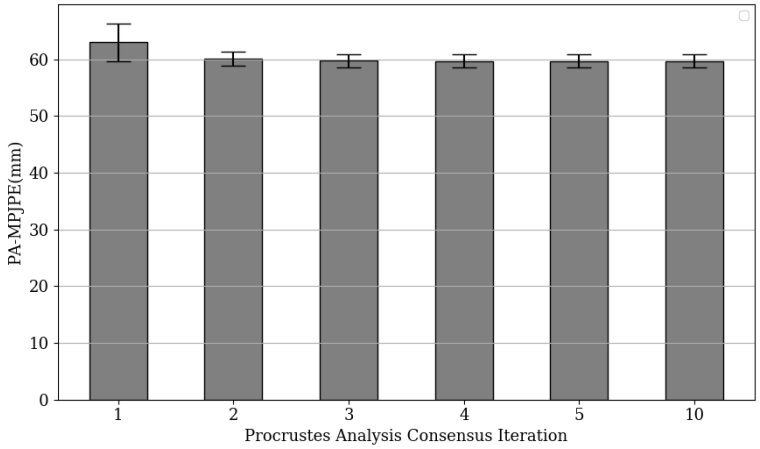

Figure 5.22. Effects of Procrustes analysis iteration count.

5.2.4. Effect of Procrustes Analysis Iteration

We have also examined the iteration of the Procrustes analysis process that produces the consensus pose. In Figure 4.3, we have proposed our multi-view reconstruction algorithm. As proposed in the algorithm, by default at the 10th line, we calculate *procrustPose* by aligning poses coming from each view to *avgPose*. In this case, we feed the final consensus pose of the previous iteration as the *avgPose* of the next iteration. In Figure 5.22, our experiment results show that reconstruction error decreases 62.97mm to 60.11mm in the second iteration. From the second iteration to the third iteration, 0.5mm marginal improvement is obtained. However, it is efficient to iterate the algorithm twice to obtain a significant improvement in the average of the reconstruction error. We perform these experiments on diagonally configured *4CAMs*. By the betterment of the further iteration, the performance difference among the combinations of diagonally configured *4CAMs* decreases. So, camera selection regarding our proposed configurations can yield close and better performance.

5.2.5. Discussion on the CMU Panoptic

Our experiments on the CMU Panoptic dataset show that wisely chosen four cameras (especially four diagonal cameras) with median filtering provide us better reconstruction quality. Two iterations in our reconstruction algorithm also yield bet-

ter reconstruction performance. Doing more than two iterations does not contribute further. The complexity of action affects the outer joints more than inner joints. Dynamic camera selection offers a significant improvement in reconstruction accuracy, but it is not easy to decide a better configuration yielding higher accuracy. Our naive-approach, body-orientation based camera selection, does not contribute to four cameras as we expect. However, for a specific single-camera configuration, we can yield similar performance to the best single-camera performance.

5.3. Experiments on the MPI-INF-3DHP

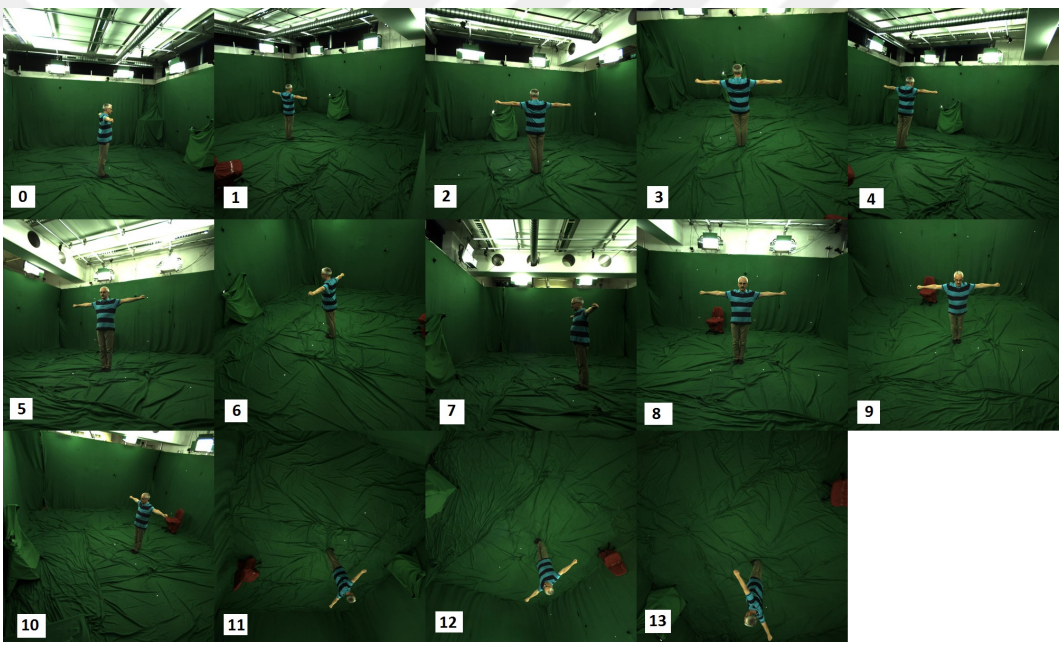

Figure 5.23. Frames from all cameras from different views in MPI-INF-3DHP.

In the MPI_INF_3DHP dataset, there are eight cameras by default. Also, there are an extra three wall cameras and three ceiling cameras available. In total, there are 14 available cameras, but we cannot utilize three ceiling cameras since our pre-trained model is not trained on ceiling cameras. In Figure 5.23, the views of all available cameras are given. The places of all utilized cameras are plotted in Figure 5.24.

In Figure 5.25, we plot single-camera reconstruction performances of each camera. Camera 11, 12, and 13 (ceiling cameras) have significantly worse performance than other cameras. Another important result of the experiment plotted in Figure 5.25 is that there is no major variation in the performance of the remaining 11 cameras placed

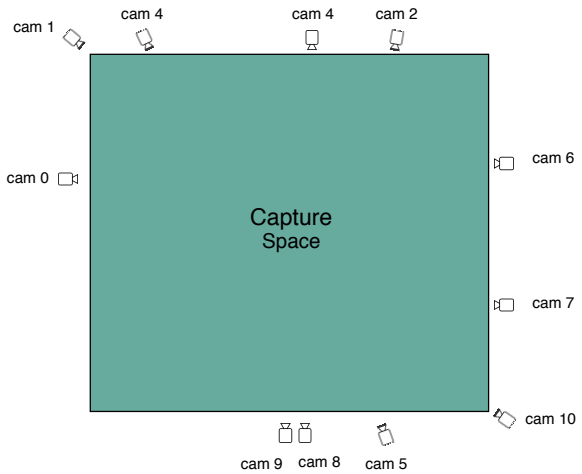

Figure 5.24. The cameras in the MPI-INF-3DHP studio from bird-eye view (except ceiling cameras).

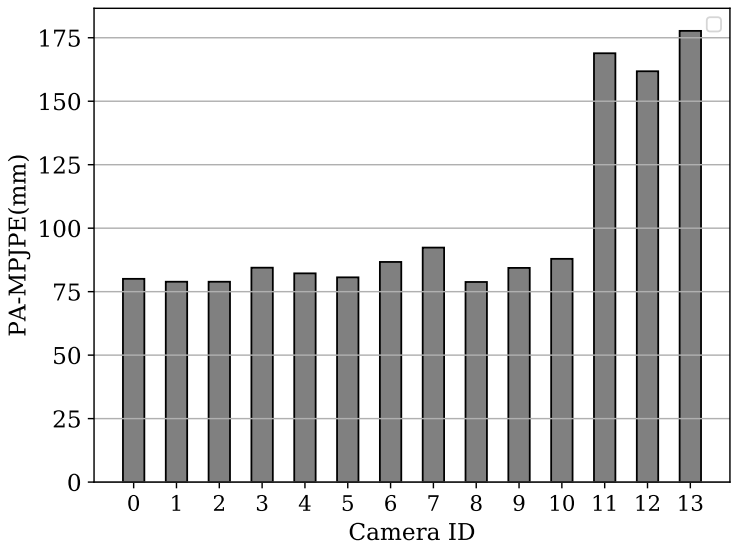

Figure 5.25. Reconstruction error for each camera on the MPI_INF_3DHP.

on walls. This is because the motion sequence in which we do our experiments spans the whole capture space. In Figure 5.26, we plot a sequence of a sample action from MPI_INF_3DHP.

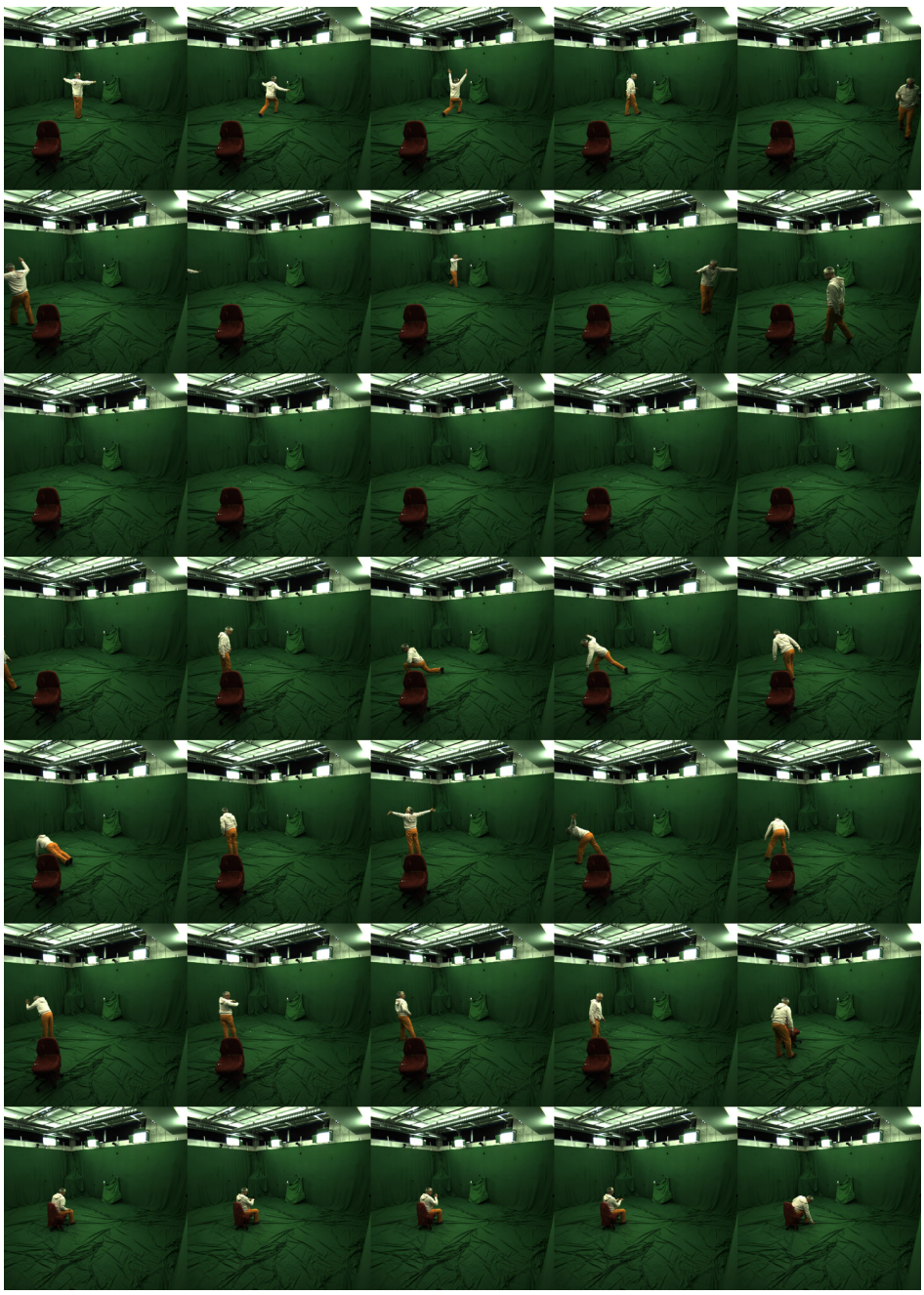

Figure 5.26. Subject 8, Video 1 of Sequence 1 (every 100th frame).

In Figure 5.27, for each case, average, minimum, and maximum error rates are plotted, except for the ALL(11) case. Increasing the number of cameras more than four in a multi-view setup does not improve the best accuracy significantly. However, using one more camera tends to decrease the average error slowly.

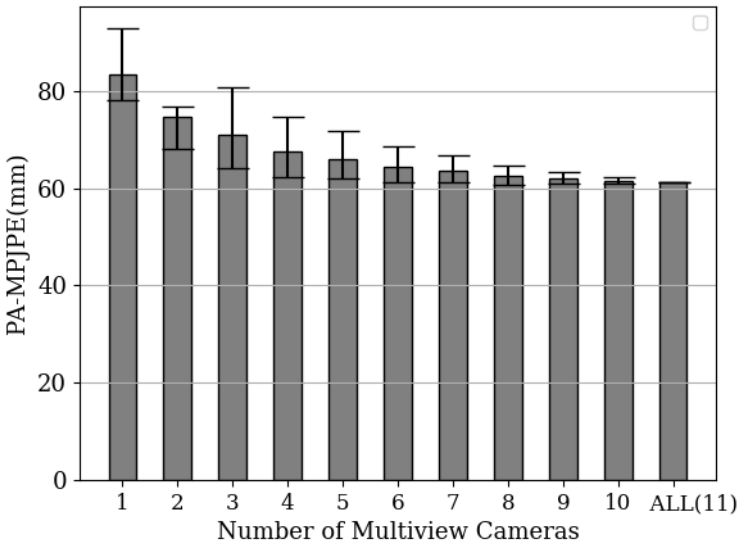

Figure 5.27. Reconstruction error for different multi-view camera setups.

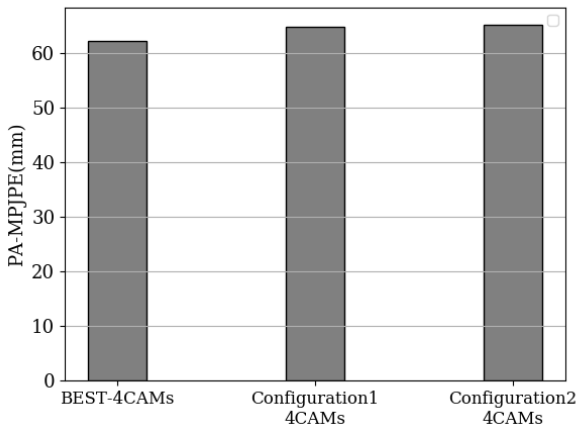

Figure 5.28. Reconstruction error (PA-MPJPE) for two different *4CAMs* configurations: *Configuration1* and *Configuration2*

As we do in our experiments on the CMU Panoptic, we compare two different four-camera configurations. In Figure 5.28, we observe that *4CAMs* with diagonal configuration show close performance to *BEST4* (impractical to know) and *ALL(11)* (costly) with a marginal loss.

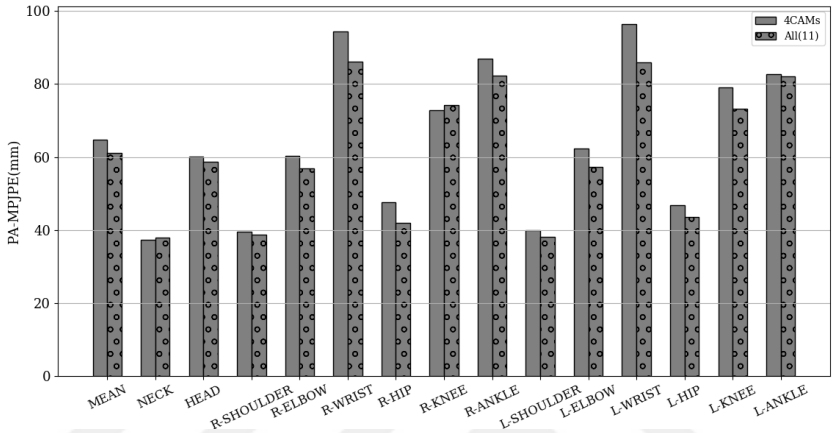

Figure 5.29. Joint-based reconstruction error analysis diagonal *4CAMs* vs *ALL(11)*.

In Figure 5.29, we evaluate joint-based performance of our proposed configuration of *4CAMs* and *ALL(11)*. The reconstruction errors in inner joints like hips, neck, etc. are similar for *4CAMs* and *ALL(11)*. Reconstruction performance even on outer joints (ankles, wrists, knees, elbows) are not significantly dominated. In avarage, *ALL(11)* is slightly better than *4CAMs*. However, the performance of wisely chosen *4CAMs* can compete with *ALL(11)*.

Table 5.1. MPI-INF-3DHP Reconstruction Error

Method	S8-Seq1	S8-Seq2	S8-Total
HMR	77.27	89.54	83.21
Ours (4CAMs)	63.77	65.78	64.74

In Table 5.1, we present the reconstruction errors on Subject 8 of the MPI_INF_3DHP dataset. While the average single-camera performance of HMR is 83.21mm of error, our reconstruction approach with four diagonal cameras with two iterations yields 64.74mm of error.

5.4. Experiments on Human3.6M

In the Human3.6M dataset, four different cameras are placed at the corners of the studio, as seen in Figure 5.30. Cameras are far enough so that the keypoints of the actor always stay in the frame.

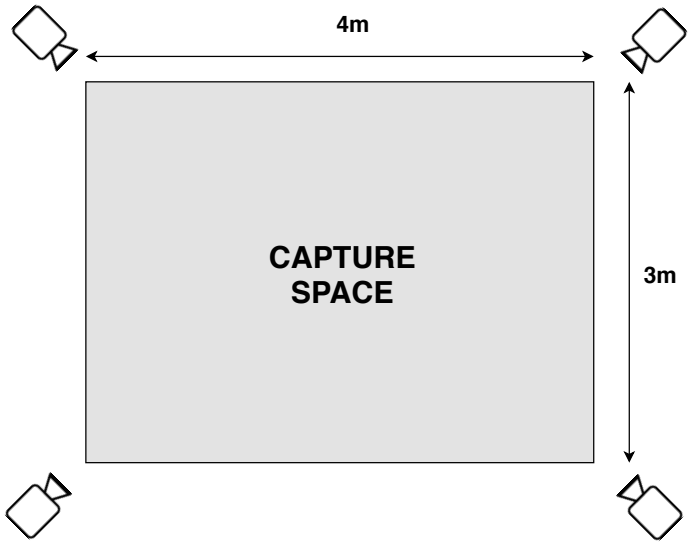

Figure 5.30. The capture area and camera setup of the Human3.6M.

The cameras are identically located in terms of elevation and angle. Moreover, all the cameras can see all human body keypoints in the frame. In the Human3.6M setup, the cameras have almost the same characteristics, so the average reconstruction errors are similar, as seen in Figure 5.31.

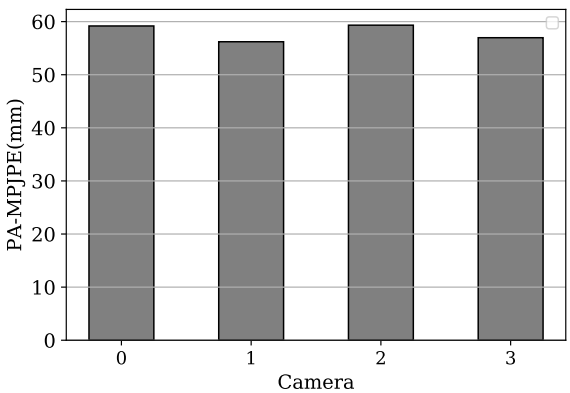

Figure 5.31. Reconstruction error for each camera on the Human3.6M.

As can be seen in Figure 5.32, some joints are more erroneous. Regarding reconstruction errors, neck, hips, and shoulders are more robust to other exterior joints

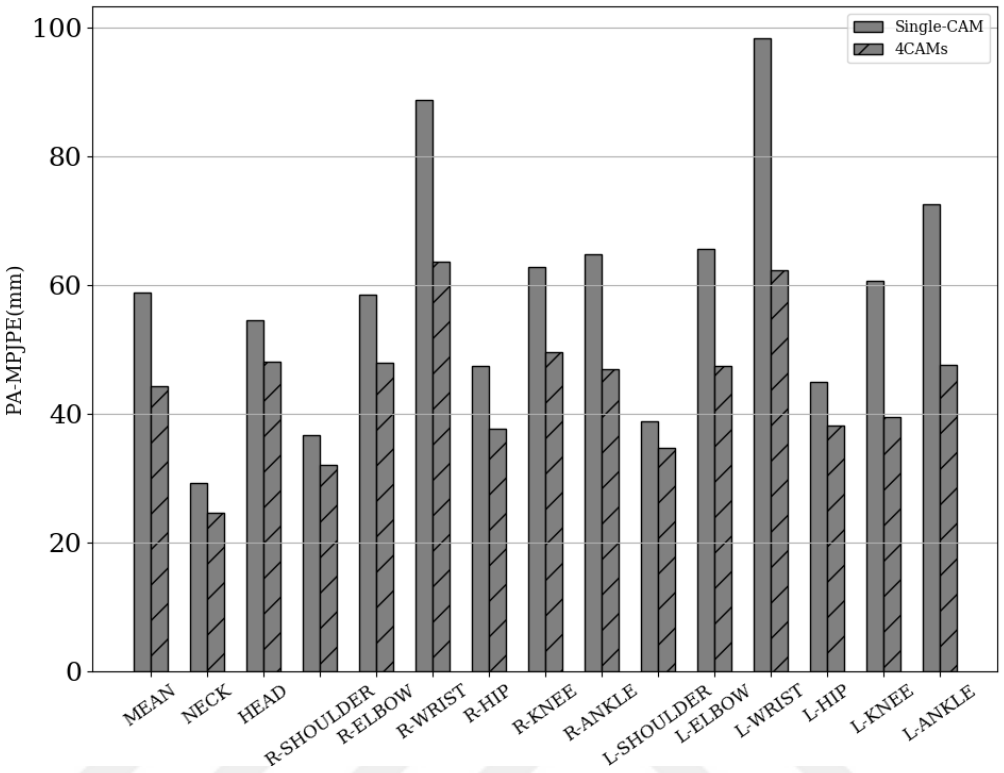

Figure 5.32. Joint-based reconstruction error analysis on the Human3.6M dataset.

like ankles. The reason why exterior joints are noisy is their higher degree of freedom. Exterior joints generally have more action-capability. The dynamicity of exterior joints also results in self-occlusion, which is another source of the reconstruction error. Another significant result is that our reconstruction improves the accuracy of exterior joints by using all four cameras.

We choose two evaluation protocols from the Human3.6M dataset: Protocol-I and Protocol-II:

- Human3.6M Protocol-I is defined such that S1, S5, S6, S7, S8 subjects are used as training, and S9 and S11 subjects are for testing. Generally, testing is performed on every 5th frame of the sequence. Error is evaluated without alignment. This benchmark not only measures the reconstruction accuracy; but also evaluates rotation and scaling correctness of the prediction technique.
- Human3.6M Protocol-II is defined by Bogo *et al.* [12]. The protocol-II is also trained on the same subject set as Protocol-I. Train/Test set splits and frame

Table 5.2. Human3.6M Results

Method	**MPJPE [Protocol-I]**	**PA-MPJPE [Protocol-II]**
HMR	86.20	57.21
Ours (4CAMs)	78.09	44.28

downsampling rates are also the same. The frontal camera is chosen from the only trial-1 of each action set. Error is calculated after the predicted skeleton is aligned with ground-truth. This error is commonly referred to as the reconstruction error.

In Table 5.2, for Protocol-I, using all four cameras shows the best performance. While single-camera performance with using the frontal camera (as in the protocol) is 86.20mm, our reconstruction method with using all four cameras yields 78.09mm absolute error (MPJPE). For Protocol-II, our method with four cameras obtains 44.28mm reconstruction error (PA-MPJPE). In Table 5.3 and Table 5.4, we compare the performance of our multi-camera reconstruction technique with the literature in the guidance of Protocol-I and Protocol-II. In Protocol-I and Protocol-II, our method is better than HMR in terms of the reconstruction performance on almost all actions.

Table 5.3. Comparison with literature on the Human3.6M dataset using Protocol-I.

Method	Direction	Discuss	Eating	Greeting	Phoning	Posing	Purchases	Sit	Sit Down	Smoking	Photo	Waiting	Walk	WalkDog	WalkTogether	Mean
Zhou [29]	87.40	109.30	87.10	103.20	116.20	106.90	99.80	124.50	199.20	107.40	143.30	118.10	79.40	114.20	97.70	113.00
Pavlakos [26]	41.18	49.19	42.79	43.44	55.62	40.33	63.68	97.56	119.90	52.12	46.91	42.68	41.79	51.93	39.37	56.89
Pavlakos [30]	48.50	54.40	54.40	52.00	59.40	49.90	52.90	65.80	71.10	56.60	65.30	52.90	44.70	60.90	47.80	56.20
Martinez [31]	51.80	56.20	58.10	59.00	69.50	55.20	58.10	74.00	94.60	62.30	78.40	59.10	49.50	59.10	52.40	62.90
HMR [7]	74.12	83.79	79.98	82.08	90.79	77.65	79.46	97.64	107.62	84.52	103.52	79.31	72.59	90.57	82.67	86.20
Ours(4Cam)	61.77	79.13	77.27	70.31	80.67	72.03	69.69	92.86	100.49	76.42	92.00	70.02	66.20	85.90	76.54	78.09

Table 5.4. Comparison with literature on the Human3.6M dataset using Protocol-II.

Method	Direction	Discuss	Eating	Greeting	Phoning	Posing	Purchases	Sit	Sit Down	Smoking	Photo	Waiting	Walk	WalkDog	WalkTogether	Mean
Zhou [29]	99.70	95.80	87.90	116.80	108.30	93.50	95.30	109.10	137.50	106.00	107.30	102.20	110.40	106.50	115.20	106.70
Bogo [12]	62.00	60.20	67.80	76.50	92.10	73.00	75.30	100.30	137.30	83.40	77.00	77.30	86.80	79.70	81.70	82.30
Pavlakos [30]	34.70	39.80	41.80	38.60	42.50	38.00	36.60	50.70	56.80	42.60	47.50	39.60	32.10	43.90	36.50	41.80
Martinez [31]	39.50	43.20	46.40	47.00	51.00	41.40	40.60	56.50	69.40	49.20	56.00	45.00	38.00	49.50	43.10	47.70
Hossain [32]	35.70	39.30	44.60	43.00	47.20	38.30	37.50	51.60	61.30	46.50	54.00	41.40	34.20	47.30	39.40	44.10
MuVs(4Cam) [33]	44.32	46.99	51.75	44.99	67.68	49.25	48.90	72.82	76.51	63.70	67.68	116.24	42.94	55.44	37.24	58.22
HMR [7]	53.22	56.75	50.41	53.94	49.37	49.37	51.39	57.78	73.70	54.39	72.90	49.97	47.09	62.64	54.95	57.21
Ours(4Cam)	37.63	42.65	38.85	43.94	45.30	40.73	39.36	49.27	60.79	45.69	51.70	40.78	35.98	48.58	42.97	44.28

5.5. Generalization to Other Methods

In this section, we evaluate the performance of our multi-view reconstruction when we use another 3D pose estimator method. Up to now, we use the HMR method as the single-view 3D pose estimator. Now, we utilize the "pose-hg-3d" method [18], which is also successful in-wild images as the HMR is. We do some experiments to check the generalizability of our multi-view reconstruction approach to other methods.

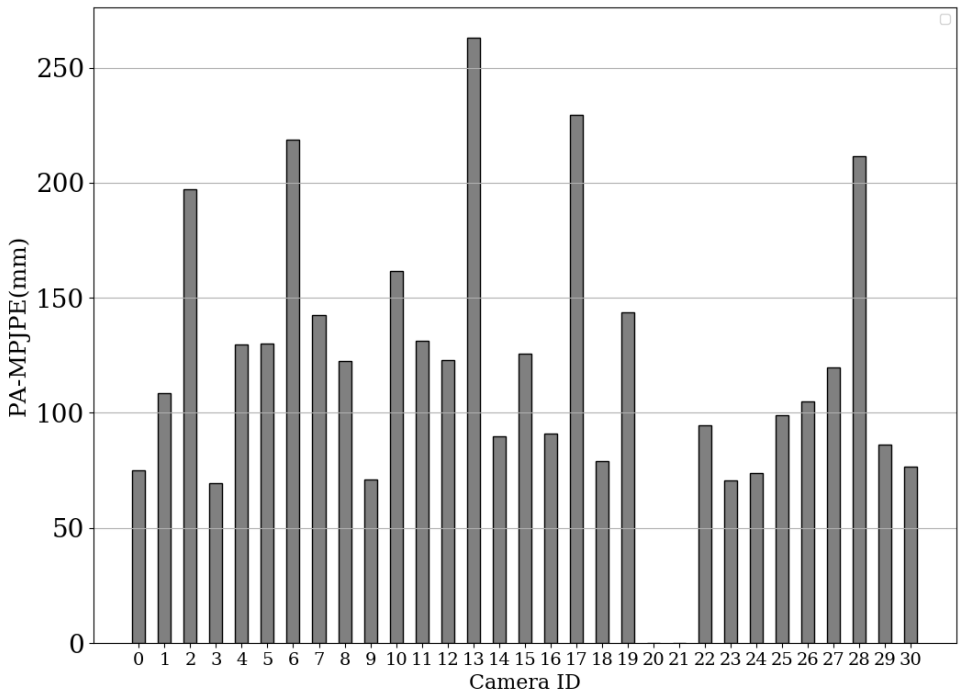

Figure 5.33. Reconstruction error (PA-MPJPE) for each single camera.

In Figure 5.33, we plot the single-camera performance of the "pose-hg-3d" method on all possible cameras on the CMU Panoptic dataset. In terms of the single-camera reconstruction performance, the "pose-hg-3d" method is worse than the HMR. Another important result is that the single-camera performance is not so similar among the remaining cameras. For HMR, the variation of the performance among remaining cameras is not so significant as in "pose-hg-3d". This shows that HMR is more robust to changes in camera view than "pose-hg-3d".

The experiments in this section show that our multi-view reconstruction approach can be combined with different single pose estimation methods. Another significant

result is that while the average single-camera performance of HMR is profoundly better than "pose-hg-3d", multi-view reconstruction performances of diagonally placed four-camera are highly similar for HMR and "pose-hg-3d" methods.

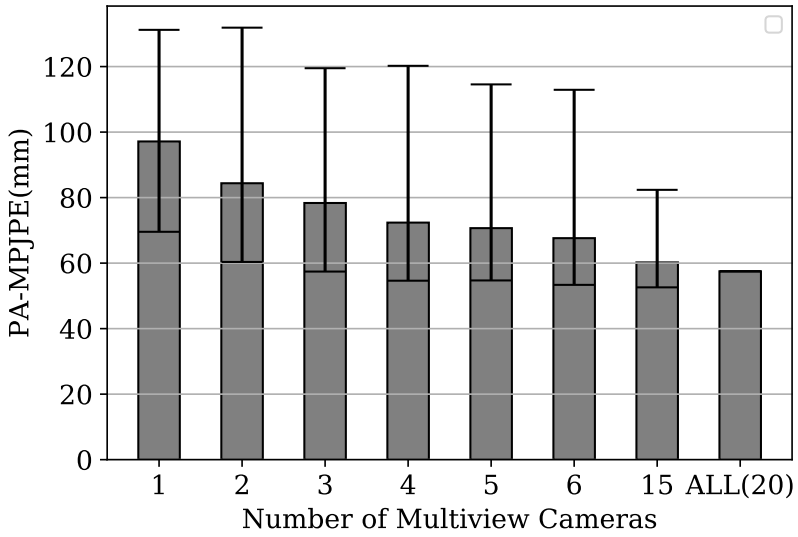

Figure 5.34. Reconstruction error (PA-MPJPE) for different multi-view camera setups.

Our experiments in this section are on the CMU Panoptic dataset. We also exclude the same upper cameras to keep the testing scenarios fixed. So, we repeat the experiments with the same remaining 20 HD cameras. As seen in Figure 5.34, while adding one more camera decreases the average reconstruction error, the best accuracy stays almost similar for the cases that more than four cameras are utilized. This result validates the conclusion of which the HMR is used as the single-view pose estimator.

In Figure 5.35, we compare the two different four-camera configurations: Configuration 1 is diagonally selected four cameras, *Configuration2* is perpendicularly selected four cameras. As seen in Figure 5.35, diagonal configuration is significantly better than perpendicular configuration. In "pose-hg-3d", the difference between two four-camera configurations is more significant since HMR is more view-invariant than "pose-hg-3d". Also, *BEST4* is better than Configuration 1, but the difference is acceptable. Another significant result is that the performance of Configuration 1 is highly similar to ALL(20). Configuration 1, diagonal four-camera, is a highly cost-effective solution when regarding the difference between ALL(20) and itself.

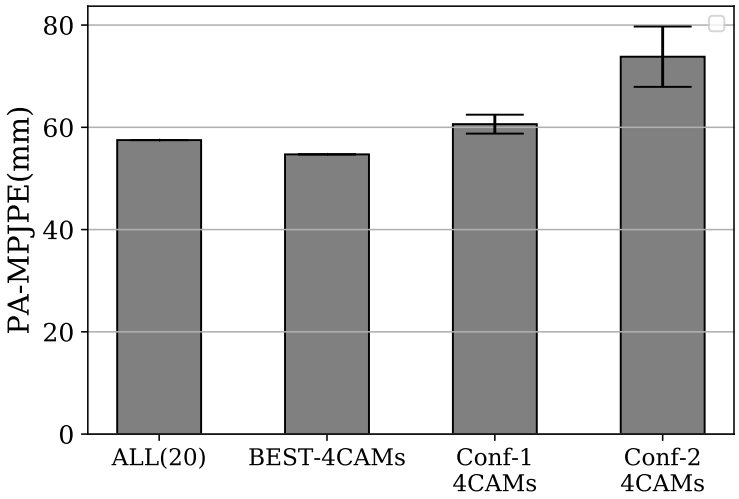

Figure 5.35. When "pose-hg-3d" is used, reconstruction error (PA-MPJPE) for two different *4CAMs* configurations: *Configuration1* and *Configuration2*.

5.6. Run Time Analysis

To examine the run time of the proposed framework, we use a computer with Intel® Core™ i7-3770 (3.40GHz × 8), 12GB RAM, GTX1070 8GB RAM GPU, and 512GB SSD space. The operating system is Ubuntu 18.04 LTS.

In Table 5.5, we plot the average execution time of our multi-view reconstruction algorithm when the predictions of the camera views are already calculated. In our physical environment, multi-view reconstruction costs 2.68ms for 20 camera views. Time cost decreases to 0.66ms for the case of four-camera views. In our testing environment, our Procrustes-based multi-view reconstruction can be calculated fast enough to be used in a real-time application.

In Table 5.6, we plot the time cost of HMR's single-view pose prediction. We can feed the batches of images to the model for 3D pose prediction. When we use all 20 cameras, we feed all 20 images as a batch to the model. In our GPU environment, this calculation takes 76.28ms in an average of 1000 times repetition. When we use four cameras, feeding four images as a batch takes 20.40ms in an average of 1000 times repetition. In this time analysis, we assume that images are given as cropped and sized

Table 5.5. Run Time Cost of Multi-view Reconstruction

# of Camera	Exec.Time(sec)
20	0.00268
15	0.00207
10	0.00144
5	0.00079
4	0.00066
3	0.00055
2	0.00041

224x224, which is the acceptable image size of the encoder part of the HMR.

Table 5.6. Run time of Single-view 3D Pose Predictions as a Batch.

# of Camera	Exec.Time(sec)	Var.of Exec.Time
20	0.076821	0.00140
4	0.020404	0.00109

As a result, if we decide to use all 20 cameras, the total time cost of one multi-view reconstruction in our testing environment is 78.96ms on average. Such performance means almost 12 predictions per second (PPS) on average. For a real-time application, 12 PPS is slightly slow in our environment. For the case of four cameras is utilized, the total time cost is 21.06ms on average. This yields almost 47 PPS, which is enough to be used in a real-time application.

6. CONCLUSION

In human pose estimation systems, techniques that utilize multi-view information are more robust to self-occlusions. If particular views cannot estimate body coordinates reliably, integration of the information obtained from other views can mitigate the performance loss. In this thesis, we showed that using a multi-view camera configuration improves the 3D pose estimation performance in terms of the reconstruction error of body joint coordinates. For this purpose, we have employed a state-of-the-art single-view human pose estimator (the HMR method) and showed that estimated body coordinates become more accurate if we use multiple camera views. In the proposed framework, individual 3D body pose information obtained from each camera view is combined with the Procrustes Analysis.

In our experiments, we have found that increasing the number of cameras results in lower reconstruction error. We have observed that carefully selected four-camera configurations offer the best trade-off in terms of reconstruction error and system complexity, i.e., the number of cameras used in a multi-view setup. Adding more cameras to the four-camera configurations yields marginal accuracy improvement. In our experiments, we have found that four cameras that are placed diagonally achieve sufficiently accurate pose estimation performance.

We have also observed that the positive impact of diagonal four-camera selection decreases when the motion sequence spans the whole studio. Another case that decreases the positive impact of diagonal camera selection is the existence of extreme poses in action. However, selecting four cameras from each quadrant still contributes more than the other four-camera configurations.

In our experiments, we have shown that *dynamic-camera selection* has a significant effect on reconstruction accuracy. However, our method, *body-orientation based dynamic-camera selection* for four cameras, did not improve the reconstruction performance. As one of our future work, we plan to design an effective way of dynamic

camera selection.

As another future work, we plan to define a Procrustes Analysis based loss function to train a ConvNet-based model on multi-view images to obtain further inference in terms of pose accuracy.

We also demonstrated that two iterations for our Procrustes Analysis based reconstruction algorithm yields lower reconstruction errors. Further iterations have presented almost the same performance. By default, our experiments on the MPI_INF_3DHP and the Human3.6M datasets are obtained by two iterations.

We have also checked the generalization of our multi-view reconstruction method to another single-view 3D pose estimator: "pose-hg-3d". Even though in terms of the single-view performance, the "pose-hg-3d" method is not better than HMR, the multi-view reconstruction performance of both two methods are highly similar in our experiments. We have also found out that "pose-hg-3d" can perform similarly in our multi-view reconstruction, even though it is not as robust to changes in camera-view as HMR is. We have observed that our reconstruction with multiple cameras can compensate for the weakness of the single-view method in camera-view dependency. Therefore, we can state that our multi-view reconstruction method can cooperate with other single-view 3D pose estimator than HMR.

We have achieved to significantly improve the performance of the HMR method in the benchmarks: Protocol I, and II. Notably, in Protocol II, our results are highly competitive in the literature.

REFERENCES

1. Ionescu, C., 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*, Vol. 36, No. 7, pp. 1325–1339, July 2014.

2. Mehta, D., H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu and C. Theobalt, "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision", *3D Vision (3DV), 2017 Fifth International Conference on*, IEEE, 2017, `http://gvv.mpi-inf.mpg.de/3dhp_dataset`.

3. Joo, H., H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara and Y. Sheikh, "Panoptic Studio: A Massively Multiview System for Social Motion Capture", *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

4. CMU Panoptic, *CMU Panoptic*, `http://domedb.perception.cs.cmu.edu/`, accessed in July 2019.

5. Jackson, A. S., C. Manafas and G. Tzimiropoulos, "3D Human Body Reconstruction from a Single Image via Volumetric Regression", L. Leal-Taixé and S. Roth (Editors), *Computer Vision – ECCV 2018 Workshops*, pp. 64–77, Springer International Publishing, Cham, 2019.

6. Rhodin, H., F. Meyer, J. Spörri, E. Müller, V. Constantin, P. Fua, I. Katircioglu and M. Salzmann, "Learning Monocular 3D Human Pose Estimation from Multi-view Images", pp. 8437–8446, June 2018.

7. Kanazawa, A., M. J. Black, D. W. Jacobs and J. Malik, "End-to-end Recovery of Human Shape and Pose", *Computer Vision and Pattern Regognition (CVPR)*, 2018.

8. The Captury, *The Captury*, `http://thecaptury.com/`, accessed in July 2019.

9. Andriluka, M., L. Pishchulin, P. Gehler and S. Bernt, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

10. Lin, T., M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft COCO: Common Objects in Context", *CoRR*, Vol. abs/1405.0312, 2014, `http://arxiv.org/abs/1405.0312`.

11. Loper, M., N. Mahmood, J. Romero, G. Pons-moll and M. J. Black, "SMPL : A Skinned Multi-Person Linear Model", *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, Vol. 34, No. 6, pp. 248:1–248:16, 2015.

12. Bogo, F., A. Kanazawa, C. Lassner, P. Gehler, J. Romero and M. J. Black, "Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image", *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, Springer International Publishing, Oct. 2016.

13. Lassner, C., J. Romero, M. Kiefel, F. Bogo, M. J. Black and P. V. Gehler, "Unite the People: Closing the Loop Between 3D and 2D Human Representations", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017.

14. Tung, H.-Y., H.-W. Tung, E. Yumer and K. Fragkiadaki, "Self-supervised Learning of Motion Capture", I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Editors), *Advances in Neural Information Processing Systems 30*, pp. 5236–5246, Curran Associates, Inc., 2017.

15. Pavlakos, G., L. Zhu, X. Zhou and K. Daniilidis, "Learning to Estimate 3D Human Pose and Shape from a Single Color Image", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 459–468, 2018.

16. Omran, M., C. Lassner, G. Pons-Moll, P. V. Gehler and B. Schiele, "Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation", *2018 International Conference on 3D Vision (3DV)*, pp. 484–494, 2018.

17. Yao, P., Z. Fang, F. Wu, Y. Feng and J. Li, "DenseBody: Directly Regressing Dense 3D Human Pose and Shape From a Single Color Image", *CoRR*, Vol. abs/1903.10153, 2019.

18. Zhou, X., Q. Huang, X. Sun, X. Xue and Y. Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach", pp. 398–407, 2017.

19. Tekin, B., A. Rozantsev, V. Lepetit and P. Fua, "Direct Prediction of 3D Body Poses from Motion Compensated Sequences", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–1000, 2016.

20. Pavlakos, G., X. Zhou, K. G. Derpanis and K. Daniilidis, "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1263–1272, July 2017.

21. Yang, W., W. Ouyang, X. Wang, J. Ren, H. Li and X. Wang, "3D Human Pose Estimation in the Wild by Adversarial Learning", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

22. Amin, S., M. Andriluka, M. Rohrbach and B. Schiele, "Multi-view Pictorial Structures for 3D Human Pose Estimation", *Procedings of the British Machine Vision Conference 2013*, pp. 45.1–45.11, 2013.

23. Belagiannis, V., S. Amin, M. Andriluka, B. Schiele, N. Navab and S. Ilic, "3D Pictorial Structures for Multiple Human Pose Estimation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 38, June 2014.

24. Belagiannis, V., S. Amin, M. Andriluka, B. Schiele, N. Navab and S. Ilic, "3D Pic-

torial Structures Revisited: Multiple Human Pose Estimation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 10, pp. 1929–1942, Oct. 2016.

25. Felzenszwalb, P. F. and D. P. Huttenlocher, "Pictorial Structures for Object Recognition", *International Journal of Computer Vision*, Vol. 61, No. 1, pp. 55–79, Jan. 2005.

26. Pavlakos, G., X. Zhou, K. G. Derpanis and K. Daniilidis, "Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

27. Varol, G., D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev and C. Schmid, "BodyNet: Volumetric Inference of 3D Human Body Shapes", *The European Conference on Computer Vision (ECCV)*, September 2018.

28. Kocabas, M., S. Karagoz and E. Akbas, "Self-Supervised Learning of 3D Human Pose using Multi-view Geometry", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

29. Zhou, X., M. Zhu, S. Leonardos, K. G. Derpanis and K. Daniilidis, "Sparseness Meets Deepness: 3D Human Pose Estimation From Monocular Video", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

30. Pavlakos, G., X. Zhou and K. Daniilidis, "Ordinal Depth Supervision for 3D Human Pose Estimation", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7307–7316, 2018.

31. Martinez, J., R. Hossain, J. Romero and J. J. Little, "A Simple Yet Effective Baseline for 3d Human Pose Estimation", *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2659–2668, 2017.

32. Hossain, M. R. I. and J. J. Little, "Exploiting Temporal Information for 3D Human

Pose Estimation", *The European Conference on Computer Vision (ECCV)*, 2018.

33. Huang, Y., F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter and M. J. Black, "Towards accurate marker-less human shape and pose estimation over time", pp. 421–430, 2017.