Chapter 4

DEEP LEARNING APPROACHES FOR 3D HUMAN POSE ESTIMATION FROM SINGLE VIEW PERSPECTIVE

This chapter describes the deep learning-based models developed for 3D HPE from single-view data input. Section 4.1 discusses the background of both models. In Section 4.2, a brief introduction of state-of-the-art methods is given. Section 4.3 provides a detailed description of the proposed models. Section 4.4 presents the result and discussion. At last, Section 4.5 concludes the chapter.

4.1 Introduction

3D Human pose estimation (HPE) is a task in which a 2D representation image is given, and its corresponding 3D pose is reconstructed. The closer the 3D pose is with the true human body actually present in the image, the better is the reconstruction. The main idea is to detect the joints of a person and form a skeleton. 3D HPE has a wide application in many fields such as human action recognition, human-robot interaction, surveillance, augmented and virtual reality, motion capture, Computer-generated imagery.

This task is found to be difficult because of arbitrary camera viewpoint, variation in human appearance, and self-occlusions. Apart from these difficulties, the main challenge is the lack of the availability of 3D Ground-truth (GT) data [14]. For 2D body joint prediction, it is possible to get easily the GT data source on a large scale. Many 3D HPE methods take benefit of this availability for 2D human body joint detection and then lift it to 3D joint coordinates [14] [16]. These techniques give satisfying result but still require to improve the reconstruction error.

We have been categorized the past techniques into two approaches: marker-based system and observation-based system. The Marker-based system requires motion capture equipment for attachment of reflective markers on the human body to estimate 3D body joint coordinate. The system has many limitations like attaching the markers to body joints is time-consuming and expensive due to the motion capture requirement. The observation-based technique utilizes the recorded videos of the human subject and breaks it into frames

for further estimation. Then the low-level and high level informations have been extracted from these recorded videos for pose estimation [10]. Directly reconstructing the 3D Human Pose from an image [11], has received great interest in the last years. Many observation-based techniques directly reconstructing the 3D pose. However, the techniques do not give satisfactory result because of less availability of 3D ground truth data. To collect the 3D ground-truth is very expensive. The deep architectures require a very large amount of training data to make the system accurate. Therefore, we follow a different workflow in both the models to address this problem, we utilize the suitability of large amount of 2D annotation data for 2D HPE. Then the generated 2D heatmaps have been incorporated for the 3D reconstruction.

Motivated by above discussion we propose two deep learning based models. In this first model, we attempt to introduce a two-stage deep neural network architecture for 3D HPE. In the first stage, a step process has been used, in the first step simple VGG-19 based CNN network used to extract features and in the second step stack-hourglass has been used to generate the 2D joint coordinates and then in the second and last stage, these generated 2D keypoints have been employed for 3D poses estimation with the help of DCDN.

In past years, many techniques have been proposed that uses only spatial data information like our first proposed model to reconstruct a 3D human pose. But many of them do not handle the projection ambiguity problem. Existing state-of-the-art techniques handle the projection ambiguity (i.e.,many 3D poses possible for a single 2D pose) by using the temporal data with recurrent deep network [21].

To address the above limitations, we propose second model with three-stage deep architecture having the workflow of 2D HPE followed by 3D HPR. This facilitates accurate 3D reconstruction. Due to the ambiguity issue of perspective projection, there are many 3D poses for a specific 2D pose. So, to solve this problem we break the 3D reconstruction into two sections: 1. Spatial 3D HPR, and 2. Temporal 3D HPR.

In the first stage, we propose an FSPE deep module for 2D HPE based on the stackhourglass technique [35] over the frames. To overcome the information loss of stackhourglass, we propose an MSCFC Strategy over the FSPE module. The resultant 2D heatmaps from the frames have been collectively utilized for 3D HPR. In the second stage, we do spatial 3D HPR with the help of basic deep learning concepts like Convolution, Batch Normalization, Dropout and RELU(Rectified Linear Units). The proposed FRC Strategy has been applied over the spatial 3D reconstruction for making it more accurate by adding the intermediate texture features. In the third and last stage, temporal 3D HPR has been done by utilizing the LSTM deep architec- ture over the generated 3D poses form the spatial stage. The temporal information has been added to make the system more accurate towards 3D pose reconstruction. The pipeline of the method is shown in Fig 4.1.

The experimental output of the proposed methods shows that the methods gives state-of-the-art performance for PCK, MPJPE, and P-MPJPE evaluation metrics.

4.2 Literature

Estimating human posture has been a crucial job from the beginning of computer vision research, and many studies have been conducted to estimate human posture in both 2D and 3D. The few below paragraphs covers both 2D and 3D techniques for estimating human posture, with a focus on CNN-based systems.

A. 2D Human Pose Estimation Early work on HPE in 2D was trained in the correlation among the body representation and body joint locations utilizing handcrafted features based on the pictorial structure [7], the model of deformed parts [189], image structure, or poselets [190].

DeepPose [108] practiced a CNN-oriented arrangement to regress the human body joints locations across many iterations. First, an initial pose is predicted using a holistic view, and the currently predicted pose is refined using a holistic view of appropriate portions of the image.

Convolutional Pose Machine (CPM) [166] is a methodical way to enhance the ability of the prediction at each level. Every level has utilized a CNN that also includes the one image as well as the confidence heatmaps of the previous levels. The result of the system is improved through joining from the prior step with the control of CNN will be compensated.

Newell et al. [111] presented a novel CNN based method where information is processed and consolidated on all scales to capture the distinct spatial relationships correlated to the

human body. They explicate how replicated top-down and bottom-up processing combined with interim monitoring is necessary for enhancing network performance. Chu et al. [9] utilized the stacked hourglass concept and incorporate the multi-context information to make the system robust towards the problem of self-occlusion.

B. 3D Human Pose Estimation Same as the 2D cases, the initial research of 3D HPE was oriented to use the low-level features like segmentation outputs or the local shape context.

Tome et al. [191] suggested a 3D HPE approach that was benefitted with multi-stage CNN architecture with probabilistic knowledge for 3D joint locations prediction. Martinez et al. [114] introduced an end-to-end deep CNN based approach to estimate 3D pose from 2D joint coordinates. Nibali et al. [113] presented an approach that is contrast to the techniques that uses 2D joint coordinates for the 3D HPE. Here they proposed the concept of 2D marginal heatmaps, that were used to estimate 3D poses to reduce the computational overhead.

Tekin et al. [105] recently proposed a structured forecasting framework that learns to estimate poses in 3D utilizing an autoencoder. Video sequences data have utilized to use the temporal information that also serves to estimate the more precise results of the 3D HPE.

Zhou et al. [107] applied the 2D HPE result to estimate the 3D poses. They designed a 3D HPE as a weighted sum of form bases that resemble a typical non-rigid construction made of motion features, and they originated an EM method that formulates the 3D pose

as a latent variable when results of the 2D HPE are also available. The system, combined with 2D HPE predictions learned from CNN, delivered the latest performance in 3D HPE.

DeepPose [45] utilized the DCNN-based composition with multiple iterations to regress the body joint locations. Firstly, by using the holistic view, it estimates initial pose and then refine these predicted poses by using important image parts.

Pishchulin et al. [31] proposed a detection-based method that uses CNN based part detection to estimate multiple people pose from images. Newell et al. [10], introduced an architecture having repeated top-down and bottom-up processing for the collection of multiple features at different scale i.e., "stacked hour-glass" for HPE from images. Belagiannis et al. [19] proposed a ConvNet based method for regressing heatmap for each body joints. This section gave a new recurrent architecture to improve the result iteratively and to make the system end to end trainable. Xiao et al. [20], introduced a method for pose estimation named simple baseline. The architecture has a deconvolutional layer over the last convolution stage of ResNet. It shows that the model is very simple as compared to other state of the art to generate heatmaps from deep and low-resolution features.

Same as 2D case, starting techniques of 3D has been based on segmentation and local shape features. These features have been utilized with structured SVM [34] and the relevance vector machine [33] for 3D HPR. Recently, Fan et al. [42] gave the summary and comparative study of the various state-of-the-art local features, utilized for the task of 3D reconstruction. Lately, after the invention of deep learning, CNN has been highly utilized

for 3D HPR. In starting, Li et al. [32] utilized the CNN for 3D HPR where the joint locations were predicted by regressing it with the learning from parent joints. Tekin et al. [44] introduced a technique that utilizes the motion information from consecutive frames for 3D HPR. Wang et al. [36] introduced a two-step method for 3D HPR that individually makes the bidirectional dependences of the body parts having varying DOFs. The first step utilized the multi-stage architecture that estimates the 2D pose from the image. The second step was progressive 3D HPR method. Zhang et al. [1] proposed a 3D HPR method by fusing the features of both 2D and 3D joints using residual learning. Katircioglu et al. [9] proposed a deep learning based regression network for reconstruction of 3D pose or 2D heatmaps. Pavlakos et al. [46] utilized the convnet for the prediction of 3D joint likelihood from the image directly.

Despite of so much progress in the field, there is huge requirement to reduce the reconstruction error. The first proposed technique aims to give a two-stage approach using deep learning concept that uses only spatial data information for 3D HPE. The second proposed technique aims to give a three-stage approach using deep learning concept that uses spatial and temporal data information for 3D HPE. These techniques indicates better result contrast to state-of-the-art techniques in respect of PCK, MPJPE and P-MPJPE evaluation metric over MPII, Human3.6M, and HumanEva-I datasets.

4.3 Methods

4.3.1 Two-stage deep network for 3D human pose estimation by exploiting spatial data via its 2D pose

The framework of the proposed approach is demonstrated in Fig. 4.1. This part of the article discusses the proposed two-stage architecture for 3D HPE. In the first stage, the two-step procedure has been given where in the first step VGG-19 based feature extraction has been used and in the second step, hourglass architecture has been used. The resulted 2D joint coordinates have been fed into the second stage of the network. The second stage utilizes the densely connected deep network (DCDN), where total of six levels have been proposed, the first level has six dense layers followed by activation function out of which three dense layers use dropout and rest of them not. The second to sixth level utilizes six dense layers followed by activation, where third dense layer uses dropout.

The main objective of the introduced network is to improve the system performance by reducing the reconstruction error. The detailed description of the both stages are discuussed in below subsections.

- **First stage of the network:** In this part, we suggest a two-step deep network where in the first step small part of a very deep VGG-19 network has been used as shown in Fig. 4.2. In the second step, four stages of the stacked-hourglass network have been utilized. The input is the RGB frame and output is 2D body joint coordinates.

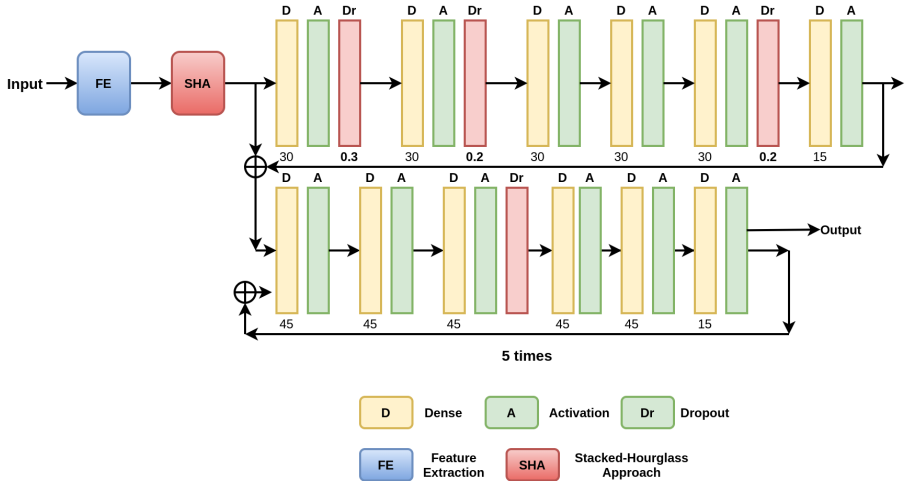

FIGURE 4.1: The framework of the proposed method.

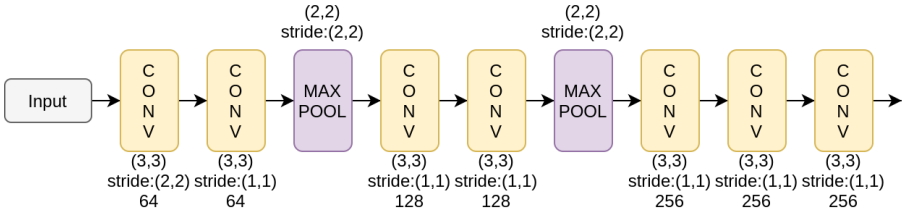

FIGURE 4.2: First stage of the network.

- **Second stage of the network:** The previous stage output is act as an input for this stage. The output is the 3D joint coordinates. Here total 6 levels of DCDN have been used. In the first level, total six dense layers have been employed followed by tanh activation function out of which three uses dropout as shown in Figure 4.1. The second to sixth level architecture is the same with six dense layers followed by tanh activation function out of which one uses the dropout layer.

4.3.2 Three stage deep network for 3D human pose estimation by exploting spatial and temporal data

In this section, we present the overall strategy of the proposed 3D HPR technique as shown in Figure 4.1. The proposed deep learning-based system comprises three sections: 2D HPE from an image named FSPE 6.5, a spatial 3D HPR technique named $Spatial_{3D}Reconstruction$ 4.3.2.2 and temporal 3D HPR named $Temporal_{3D}Reconstruction$ 6.3.2.1. We have named $STPR_{3D}$ 3 to the 3D reconstruction method created from using both the second and third stages. The FSPE method has been applied over the video frames. We observe that the adjacent sequences of a particular frame are mostly correlative. So, we found that, more than one frame consideration gives more natural and smooth reconstruction. The resulted 2D heatmaps have been used for 3D reconstruction using $Spatial_{3D}Reconstruction$, where the basic deep learning concepts like Convolution, Batch Normalization, Dropout and ReLU have been utilized. At last, the temporal features have been used with the LSTM model for final reconstruction. The detailed description of the architecture is in the following sections:

4.3.2.1 Frame specific pose estimation(FSPE) 2

The network proposes a 2D HPE method for an image. The method has the main two stages of feature extraction and feature refinement. The feature extraction method is a multi-level CNN based method inspired by the inception-v4 [167] and VGG-19 [168]

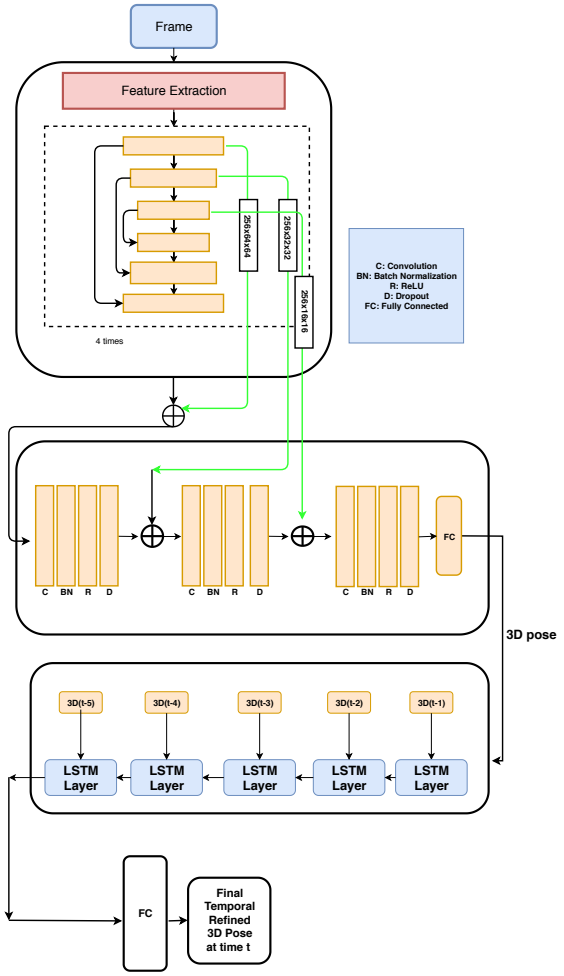

FIGURE 4.3: The overall pipeline of the proposed method

deep network as shown in Figure 4.2. We use the stem part of the inception-v4 for first-level feature extraction. The inception network is extremely tunable, where we can easily make modifications at the level of filter count on various layers, which does not disturb the fully trained system performance and extract relevant features which make the prediction accurate with low computational cost. The second level CNN module is a very small portion of very deep VGG. We observe that the multi-level fused CNN module for feature extraction have abundant quality knowledge to make an accurate 2D prediction. The extracted features have been fed to the feature refinement module motivated by the

stackhourglass technique. Pose estimation is more challenging than image classification. After the stack-hourglass introduced, many of the techniques in literature utilize the multi-level network for pose estimation. So, we utilize four stages of the hourglass for feature refinement in our module. We observe that the stack-hourglass technique has been insufficient for pose estimation because of its design choice, they use repetitive up and down sampling, which leads to huge chances of information loss. We work on this limitation and propose an MSCFC, over every stage of the feature refinement with a total of four hourglass stages. The MSCFC strategy circulates the multiple scale information, which alleviates the training obstruction and reinforces the data flow. The MSCFC connections

Algorithm 2: FSPE(Frame Specific Pose Estimation)

Input: Frame
Output: 2D Pose
Procedure
Feature Extraction:
$Input \leftarrow image : x$
$x_1 \leftarrow First - level_{FE}(x)$
$x_2 \leftarrow Second - level_{FE}(x)$
$x \leftarrow add(x_1, x_2)$
Feature Refinement(x):
MSCFC(Multi-stage cascaded feature connection approach):
stage=4
for $i \leftarrow 1$ **to** $stage - 1$ **do**
    d_1, d_2, d_3**: Current stage down sampling output of the hourglass**
    c_1, c_2, c_3**: Current stage up sampling output of the hourglass**
    $d_1^{'}, d_2^{'}, d_3^{'}$**: next stage down sampling input of the hourglass**
    $d_1{}' \leftarrow$ **add**(d_1, c_1)
    $d_2{}' \leftarrow$ **add**(d_2, c_2)
    $d_3{}' \leftarrow$ **add**(d_3, c_3)
    Return$\leftarrow$ **laststage**(c_1)
end
Return $\leftarrow$ **joint-heatmaps**

over the feature refinement module are shown in Figure 4.3 with a green color of connection arrow. For every scale, two separate data flows are proposed from downsampling and upsampling units of the preceding stage to the downsampling operation of the present stage. The 1×1 convolution operation have been added on every flow. The network outputs the joint heatmaps. Each output of 2D pose prediction has been depicted as K heatmaps, which is the number of joints. The $y \in R^{W \times H \times 3}$ is the RGB input frame. fm_s is the midway s^{th} feature heatmap of a frame and $h_k \in R^{W_h \times H_h \times L}$ (k= 1,2,....,K)(W, H is the dimesion of the heatmap and L is the number of channels) is the K^{th} joint heatmap from a frame.

The FSPE(f) is given below for a frame as:

$$f(y) = ((h_1, ..., h_k)(fm_1, ..., fm_s)). \tag{4.1}$$

The heatmap loss for intermediate supervision is:

$$Loss_{2D} = \frac{1}{K}(\sum_{k=1}^{K} \left\| h_k - \hat{h_k} \right\|). \tag{4.2}$$

Here $\hat{h_k}$ is the ground-truth and h_k is the predicted one. The FSPE has been applied over each of the frames of the video and the resulted heatmaps have been utilized for 3D pose reconstruction described in further sections.

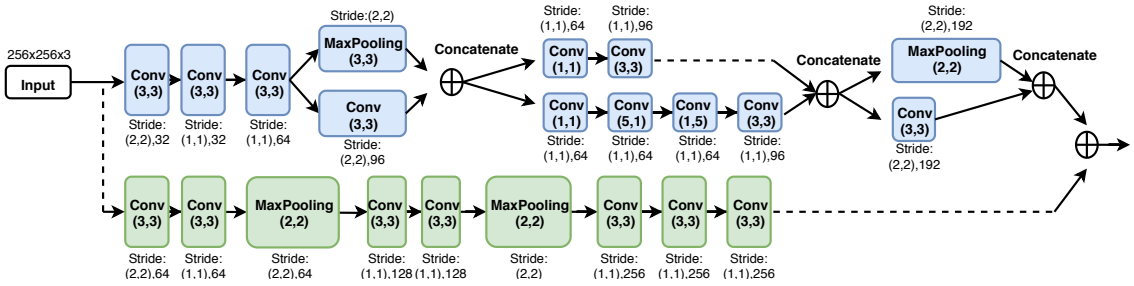

FIGURE 4.4: Feature Extraction for FSPE Module.

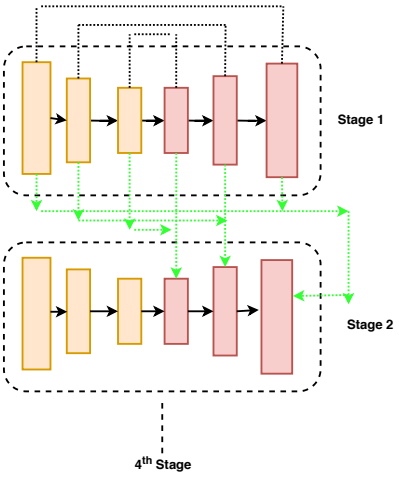

FIGURE 4.5: Feature Refinement with MSCFC Strategy.

4.3.2.2 $Spatial_{3D}Reconstruction$

The proposed technique as shown in Figure 4.1, is based on very simple deep learning concepts using Batch normalization, Dropout, Rectified Linear Units(RELUs) with Convolution. We choose to use the joint heatmap rather than joint coordinate, in contrast to recent methods [142] to reduce the extra computation. All the output 2D heatmaps are concatenated with some intermediate features using feature residual connection strategy as discussed in the below section and act as input for the proposed 3D module. The combined input has been passed through the convolutional layer for the learning of translation-invariant filters from the joint heatmaps. Batch Normalization and Dropout are utilized to

make the system more efficient by improving its performance in the cases where we train the system on the 2D estimation output. The proposed deep module has non-linearity by using RELU.

The input to this module is the concatenation of the frame specific pose heatmaps and some intermediate features using FRC strategy 6.4.2. The output is 3D pose or collection of joint coordinates in 3D space, depicted as $p \in R^{3\times k}$. The generalized form of the integrated module is as follows:

$$\hat{p} = network_{integrate}(IN_{3D}(join(h_1, ..., h_k), join(fm_1, ..., fm_s))). \quad (4.3)$$

The loss for supervision is:

$$Loss_{3D} = \frac{1}{K}(\sum_{k=1}^{K} \|p_k - \hat{p_k}\|). \quad (4.4)$$

where the p_k and $\hat{p_k}$ is the ground-truth and estimated joint coordinate for joint k.

4.3.2.3 Feature Residual connection(FRC) strategy

The ambiguity because of the generation of multiple 3D poses for a specific 2D image is a major issue of 3D reconstruction. We proposed a Feature residual connection strategy to overcome the issue. The method utilizes the intermediate features along with the generated 2D heatmap for input to the 3D pose reconstruction [192]. With the help of

Algorithm 3: $STPR_{3D}$**(Spatial Temporal Pose Reconstruction in 3D Space)**

Input: 2D Heatmaps

Output: 3D Pose

Initialize: let NS is the number of video sequences, k is the number of joints, h is the heatmap, conv is convolution, BN is Batch Normalization, R is ReLU and D is Dropout.

Procedure for section 4.3.2.2:

`Spatial(`x`)`

x← conv(x)

x← BN(x)

x← R(x)

x← D(x)

return x;

x ← **add(**IN_{3d}**(join(**h_1^1**,.....,**h_1^{NS}**),....,join(**h_k^1**,....,**h_k^{NS}**),**d_1**):x**

x'**=Spatial(x)**

x'**=Spatial(**x'**,**d_2**)**

x''**=add(**x''**,**d_3**)**

x''**=Spatial(**x''**)**

$\hat{p}$ **= FC(**x''**)**

Return ← **3D-Keypoints(**$\hat{p}$**)**

Procedure for section 6.3.2.1:

n is six here

$\hat{p}_1^{'}$**=**$LSTM_1(p_1)$

$\hat{p}_2^{'}$**=**$LSTM_2(\hat{p}_2, \hat{p}_1^{'})$

$\hat{p}_3^{'}$**=**$LSTM_3(\hat{p}_3, \hat{p}_2^{'})$

$\vdots$

$\hat{p}_{n-1}^{'}$**=**$LSTM_{n-1}(\hat{p}_{n-1}, \hat{p}_{n-2}^{'})$

$\hat{p}_{3D}$**=FC(**$\hat{p}_{n-1}^{'}$**)**

Return ← **Final-3D-Keypoints(**$\hat{p}$**)**

this intermediate feature, system preserves the important texture features that handles the ambiguity issue. The last stage encoder layers output of the hourglass module has been added with the residual connection to the final output shown with green arrows in Figure 4.5.

4.3.2.4 $Temporal_{3D}Reconstruction$

In the previous phase of the 3D HPR, we only considered the spatial information into account, where the pose reconstruction was carried out from the frames, without taking the pose of the preceding time step. Once the 3D pose obtained from the video frames, the temporal information has been utilized to make the system more accurate by integrating the poses with time series. The system takes advantage of historical data to process. This creates a close connection in successive time steps. The pose changing variation between successive images have been anticipated to be small. The joints and limbs trajectory should be taken smooth here.

The utilization of spatial and temporal data is motivated to improve the result. At the same time, it is anticipated that the result get less effected because of the miss detection. The module has been trained to reconstruct the pose at particular time step *t*. We have 3D poses at time step t-NSL to t-1 as input, here NSL is five. For this module, the training data is maintained as (fv_t,gt_t), where fv_t is the feature vector has a collection of 3D keypoints and gt_t is the ground-truth of the method. Here the input and output size is NSL$\times k\times$3 and $k \times 3$. The network configuration having an LSTM having 256 units of the hidden layer, come behind the FC layer having RELU activation function with $k \times 3$ hidden units. Mean Square Error function is used to compute the loss.

4.4 Experiments and Results

4.4.1 Two-stage deep network for 3D human pose estimation by exploiting spatial data via its 2D pose

The effectiveness of the given approach was checked on the HumanEva-I, Human3.6M, and MPII datasets with the use of PCK, MPJPE and P-MPJPE evaluation metrics. Here MPII and PCK are used to evaluate the 2D HPE performance. Similarly, Human3.6M, HumanEva-I, MPJPE, and P-MPJPE are used to evaluate the 3D HPR performance.

4.4.1.1 Dataset and Metrics:

MPII Dataset: MPII [127] is the dataset used for 2D HPE. This dataset contains the images as the frames utilizing the Youtube videos data that incorporate every day human activity. This includes a whole of twenty-five thousand images simultaneously with forty thousand annotations. From these thirty thousand images are appropriate to train and ten thousand to test.

Human3.6M dataset: This dataset [193] is a newly issued data collection that gives a total of 3.6 million frames for 3D HPR and identical annotations formed in controlled lab conditions. Here the dataset uses eleven acknowledged actors who perform for fifteen situations under four distinct views. Furthermore, the mostly used data partition way in literature is:

- Evaluation criteria: Here the data have collected from five different subjects (S8, S7, S6, S5, and S1) are utilized to train the system, and similarly the S11 and S9 two subjects used to test. To expand the size of the training set, the frame sequences which are taken from independent perspectives of the corresponding subject are handled as distinct sequences.

HumanEva-I dataset: HumanEva-I [194] is an enormously smaller benchmark dataset, which contains 3 subjects that are photographed from 3 views-points. The authors assess the performance on 3 actions (Box, Jog, Walk) to train the one model collectively for all actions. We utilize the provided test/train sets.

Evaluation Metric: 2D Case:

- PCKh Metric: PCK is termed as a Percentage of correct keypoints metric which is employed to predict the 2D body joint locations. The estimation is resulted true on the basis of the distance among the GT and predicted keypoints. The estimated joint locations are considered valid if the distance within GT and predicted keypoint is more inferior than the α portion of the head segment height (PCKh). This whole concept is termed as a PCKhα.

3D Case:

- Metric 1: it is the mean Euclidean distance within the GT and predicted output and it is termed as mean per-joint position error (MPJPE), measured in millimeters as

[].

$$MPJPE(f, \hat{f}) = \frac{1}{L} \sum_{l=1}^{L} \left\| e_l(f) - e_l(\hat{f}) \right\| \tag{4.5}$$

f is the GT pose

$\hat{f}$ is the estimated pose

$e_l()$ denotes the 3D location of l^{th}joint

- Metric 2: it describes the error with the alignment of GT in scale, rotation, and translation. This metric is named as P-MPJPE error.

4.4.1.2 Implementation Details:

To train and test the proposed method we have used the NVIDIA Tesla k40c of 12GB system.

4.4.1.3 Training Details

The introduced network consists of two-stage. In the earliest stage, data was resized and cropped to a size of 256x256 with the subject centered in the image of the MPII dataset. These data also augmented with random rotation(+- 40 degree), data translation (accompained by +-2% of input image size), and scaling of 0.8 to 1.2 on MPII. For the aforementioned practice, the 1e-4 is the learning rate. That falls with 0.4 to saturates the loss over the validation collection. For the aforementioned practice, the 1e-4 is the learning rate. That falls with 0.4 to saturates the loss over the validation collection. We have applied the

validation split criteria the identical as [114]. For the Human3.6M dataset, we choose an exponentially learning rate of 4e-5 with 70 epochs along with an α = 0.95 (shrink factor) employed for each epoch. For optimization purpose Amsgrd optimizer has been applied here. For the HumanEva-I dataset, we choose an exponentially learning rate of 4e-3 with 100 epochs along with an α = 0.99 (shrink factor) employed for each epoch.

4.4.1.4 Results and Ablation Analysis:

This part of the article discusses the effectiveness of the proposed method by analysing its performance over HumanEva-I, MPII, and Human3.6M datasets on PCK, MPJPE and P-MPJPE metrics.

4.4.1.5 MPII Dataset:

The performance of the proposed 2D HPE method has been checked using the evaluation protocol called PCKh metric. The method also compared with the state-of-the-arts shown in Table 4.1, like Bulat et al. [149], Newell et al. [111], Wei et al.[166], Tompson et al. [169], Carreira et al. [183], and Belagiannis et al. [158]. The result using PCK metric shows that the proposed method gives state-of-the-art performance by beating all the above mentioned methods.

4.4.1.6 Human3.6M Dataset:

The performance of the proposed 3D HPR method has been checked using the evaluation protocol called MPJPE and P-MPJPE metric. The proposed method also compared with the state-of-the-arts shown in Table 4.2 and Table 4.3. The method has been compared with Martinez et al. [114], Yang et al. [13], Huang et al. [195], Sun et al. [162], Chen et al. [196], and Pavlakos et al. [11] on MPJPE metric. The result using MPJPE metric shows that the proposed method gives state-of-the-art performance by beating all the above mentioned methods. Similarly, the performance also measured using P-MPJPE metric as shown in Table 4.3. The state-of-the-art techniques reported in Table 4.3 for comparison are Martinez et al. [114], Yang et al. [13], Sun et al. [162], Hossain et al. [197], and Fang et al. [198]. The result shows that the proposed method gives competitive performance compared to them.

4.4.1.7 HumanEva-I

The performance of the proposed 3D HPR method has been checked using the evaluation metric called MPJPE. The given method also compared with the state-of-the-arts shown in Table 4.4. The comparison has been done on the Lee et al. [199], Pavlakos et al. [11], Pavlakos et al. [200], and Martinez et al. [114]. The result shows that the suggested approach shows competitive performance compared to them.

TABLE 4.1: The outcome of proposed approach and its comparison with published state-of-the-art techniques on the basis of PCK0.5 metric.

MPII Human Pose								
Method	Ankle	Hip	Elbow	Head	Knee	Wrist	Shoulder	Total
Bulat et al. [149]	81.7	89.4	89.9	97.9	85.7	85.3	95.1	89.7
Newell et al. [111]	83.6	90.1	91.2	98.2	87.4	87.1	96.3	90.9
Wei et al. [166]	79.4	88.4	88.7	97.8	82.8	84.0	95.0	88.5
Tompson et al. [169]	64.8	80.9	83.9	96.1	72.3	77.8	91.9	82.0
Carreira et al. [183]	66.4	82.8	81.7	95.7	73.2	72.4	91.7	81.3
Belagiannis et al. [158]	78.4	87.9	88.2	97.7	82.6	83.0	95.0	88.1
Ours	84.1	89.5	92.9	98.5	87.2	87.8	97.1	91.0

TABLE 4.2: The outcome of proposed approach and its comparison with published state-of-the-art techniques on the basis of MPJPE metric.

Human3.6M								
Method	Eat	Photo	Sitting	Wait	WalkT	Discuss	Greet	Purch.
Martinez et al. [114]	58.1	78.4	74.0	59.1	52.4	56.2	59.0	58.1
Yang et al. [13]	50.4	65.4	69.2	58.4	47.7	58.9	57.0	52.7
Huang et al. [195]	22.5	40.1	26.0	22.9	21.3	20.7	24.5	23.1
Sun et al. [162]	45.0	37.6	71.4	41.6	36.9	41.4	45.2	52.0
Chen et al. [196]	40.5	34.9	67.5	37.5	34.2	39.3	41.2	51.2
Pavlakos et al. [11]	66.7	77.0	83.7	65.8	63.2	71.9	69.1	68.3
Ours	46.1	36.5	70.7	40.5	35.3	40.9	44.5	52.3
Method	Sitt.D	Direct	Phone	Pose	Walk	Smoke	WalkD	Avg
Martinez et al. [114]	94.6	51.8	69.5	55.2	49.5	62.3	65.1	62.9
Yang et al. [13]	85.2	51.5	62.1	49.8	60.1	57.4	43.6	58.6
Huang et al. [195]	39.9	18.7	28.3	22.7	22.9	33.8	35.0	26.9
Sun et al. [162]	42.5	40.9	42.1	41.1	42.6	47.4	32.0	44.1
Chen et al. [196]	42.1	36.9	42.0	38.0	40.2	42.5	30.6	41.6
Pavlakos et al. [11]	96.5	67.4	72.0	65.0	59.1	71.7	74.9	71.9
Ours	42.1	39.9	42.0	40.8	41.2	45.7	30.5	43.2

TABLE 4.3: The outcome of proposed approach and its comparison with published state-of-the-art techniques on the basis of P-MPJPE metric.

Human3.6M (P-MPJPE)								
Method	Eat	Photo	Sitting	Wait	WalkT	Discuss	Greet	Purch.
Martinez et al. [114]	46.4	56.0	56.5	45.0	43.1	43.2	47.0	40.6
Yang et al. [13]	36.3	47.4	36.9	30.5	32.2	30.9	39.9	29.4
Sun et al. [162]	45.0	53.0	59.3	44.0	44.8	44.3	45.4	41.3
Hossain et al. [197]	44.6	54.0	51.6	41.4	39.4	39.3	43.0	37.5
Fang et al. [198]	43.7	55.3	54.5	44.3	41.7	41.7	44.9	38.2
Ours	36.0	43.7	51.9	29.9	40.5	41.1	42.3	30.6
Method	SittingD	Direct	Phone	Pose	Walk	Smoke	WalkD	Avg.
Martinez et al. [114]	69.4	39.5	51.0	41.4	38.0	49.2	49.5	47.7
Yang et al. [13]	58.4	26.9	43.9	28.8	42.5	41.5	29.5	37.7
Sun et al. [162]	73.3	42.1	51.5	43.2	38.3	51.0	48.0	48.3
Hossain et al. [197]	61.3	35.7	47.2	38.3	34.2	46.5	47.3	44.1
Fang et al. [198]	64.4	38.2	48.5	40.2	36.7	47.2	47.3	45.7
Ours	59.0	26.0	43.5	27.9	33.5	40.3	28.6	38.3

TABLE 4.4: The result of proposed method and comparison with other state-of-the-art using MPJPE metric on HumanEva-I dataset.

HumanEva-I									
Method	Jog			Walk			Box		
	S1	S3	S2	S1	S3	S2	S1	S3	S2
Lee et al. [199]	25.7	17.7	16.8	18.6	30.5	19.9	42.8	53.4	48.1
Pavlakos et al.[11]	28.9	23.8	21.9	22.3	29.7	19.5	-	-	-
Pavlakos et al. [200]	23.5	14.5	15.4	18.8	29.2	12.7	-	-	-
Martinez et al. [114]	26.9	18.6	18.2	19.7	46.8	17.4	-	-	-
Ours	23.0	13.9	15.5	18.5	28.7	12.2	23.8	34.3	31.5

4.4.2 Three stage deep network for 3D human pose estimation by exploting spatial and temporal data

4.4.2.1 Dataset and evaluation protocols for 2D pose estimation

The evaluation of 2D is done on the publically available MPII dataset.

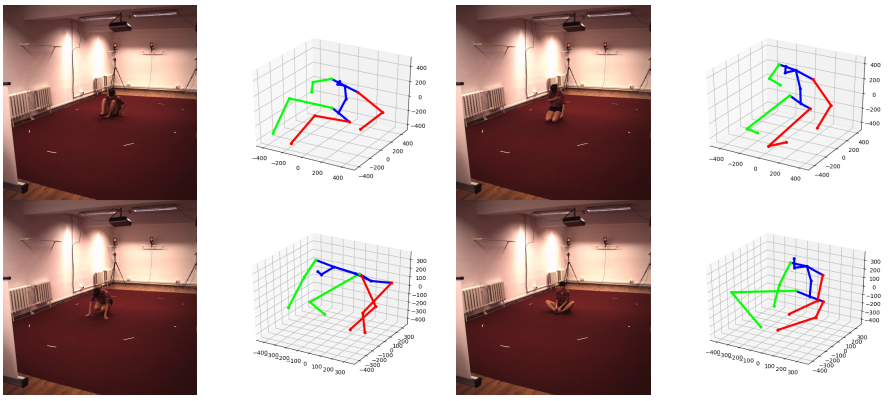

FIGURE 4.6: Visualization of the image and its 3D reconstruction.

MPII Dataset. MPII [35] is the benchmark dataset for the single person HPE. The dataset has a collection of images from the Youtube videos which contain the daily human activities. It contains a total of 25k images along with 40k annotations. The 30k images have been utilized for training and 10k for testing. Concerning other datasets, MPII gives more knowledge like fully annotated frames and activity labels with good image resolution. The keypoint locations have been utilized for training.

Evaluation Protocol(PCKh) First, we discuss PCK, where the predicted joint is treated accurately if the distance between the ground truth and predicted lies within the set threshold. Here the threshold is 0.5 and it is represented as PCK0.5. Another variant of PCK is PCKh, where the predicted keypoint is taken as correct if the distance between predicted and ground truth keypoint is lesser than α factor of head segment length (PCKh). The representation of the metric is as PCKhα.

TABLE 4.5: The result of FSPE method and comparison with other state-of-the-art using PCKh with @0.5.

Approaches	Hip	Head	Wrist	Shoulder	Elbow	Ankle	Knee	Total
Bulat et al.[149]	89.4	97.9	85.3	95.1	89.9	81.7	85.7	89.7
Gkioxary et al. [201]	85.2	96.2	82.1	93.1	86.7	74.1	81.4	86.1
Rafi et al. [185]	86.8	97.2	81.3	93.9	86.4	73.4	80.6	86.3
Wei et al. [166]	88.4	97.8	84.0	95.0	88.7	78.0	83.4	88.5
Chen et al. [186]	90.2	98.5	88.5	96.5	92.5	86.0	89.6	91.9
Belagiannis et al. [158]	87.9	97.7	83.0	95.0	88.2	78.4	82.6	88.1
Chu et al. [9]	90.6	98.5	88.1	96.3	91.9	85.0	88.0	91.5
Chou et al. [112]	91.3	98.2	88.0	96.8	92.2	84.9	89.1	91.8
Ours	92.6	99.3	88.5	98.8	93.6	85.6	90.1	92.7

4.4.2.2 Dataset and evaluation protocols for 3D pose estimation

The evaluation of the method for 3D is done on the publicaly available Human3.6M, HumanEva-I datasets.

Human3.6M Dataset. The dataset is currently introduced, having 3.6 million 3D human pose video frames and the respective annotation in a controlled environment. The dataset acquires 11 subjects with 15 actions from 4 distinct viewpoints like eating, walking, making a phone call, sitting. The evaluation criteria are the same as the few previous works [142]. The training has been performed on five subjects (S_1, S_5, S_6, S_7, S_8) and testing on 2 subjects (S_9, S_{11}). To improve the training, the various viewpoints of the same subject is taken as a different sequence for the dataset. Here one model is utilized to train for all activity.

HumanEva-I. The dataset has been generated in a controlled environment, having three standard RGB views at 60 Hz. The dataset is much smaller in size, acquire three

subjects with sic actions. We train single model for all actions. The given train/test split has been utilized here.

Evaluation Protocols. The three standard criteria have been utilized to evaluate the 3D pose:

- **Criteria 1 (MPJPE):** is the Mean per-joint position error (MPJPE), which is higlly utilized metric for 3D HPR.

  PJPE (Per-joint position error) is defined as the Euclidean distance between the predicted joint and the groundtruth.

  Mean of PJPE is the per joint error for all respective joints which is N here is given as:

$$MPJPE(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \|m_i(x) - m_i(\hat{x})\| . \tag{4.6}$$

  x: represent the ground truth pose.

  $\hat{x}$: Estimated pose.

  $m_i(x)$ represent the 3D position of the i^{th} joint.

- **Criteria 2 (P-MPJPE):** is also called reconstruction error, which is after rigid alignment of the joints with groundtruth using Procrustes Analysis. Here, the alignment perform with respect of all translation, scale and rotation, before calculating the MPJPE.

- **Criteria 3 (N-MPJPE):**here, we align the poses with the groundtruth via least-square method, respective of only scale, before calculating the MPJPE.

These metric removes the misalignment and make 3D skeleton more qualitative. The formula for error metric is computation relative to the root node between the groundtruth and estimated data is as follows:

$$P - MPJPE = (1/T)(1/N) \sum_{t=1}^{T} \sum_{i=1}^{N} \left\| (J_i^t - J_{root}^t) - (J_i^t - J_{root}^t) \right\|^2 . \tag{4.7}$$

4.4.2.3 Implementation Details:

We propose three stage training details. The testing and training has been carried out on the system with configuration of 12GB NVIDIA Tesla K40c. The time required for test is total of all the three parts:

1. Time for 2D is 0.233(s) for each video sequence.
2. Time for 3D reconstruction is 0.004(s).
3. Time for LSTM based refinement takes 0.00047(s).

Training Details. We propose a three-stage training network. In the first stage, the input image was cropped and resized to 256x256 pixels taking the main subject at the center for the MPII dataset. The data augmentation comprises random data translation (with +-2% of the given image size), random rotation (+- 40 degree) and scaling from 0.7

to 1.3 on MPII. In this experiment, the base learning rate is 1e-3. It drops by a factor of 0.4 when the loss on the validation set saturates. The used validation split is the same as [202]. The Human3.6M dataset utilized for training of 2^{nd} and 3^{rd} phase of the 3D HPR. For the second stage, the learning rate is 4e-4 with 60 epochs and drop-down a factor of 0.95. The Amsgrd optimizer has been utilized here. For the third stage, the learning rate is 1e-4 with 90 epochs and a drop-down factor of 0.98. Adam optimizer utilized here. The train/test split is the same as provided by the dataset of both 2D and 3D cases.

4.4.2.4 Result and Discussion

This section discuss about the results obtained by the proposed method. We also do the comparison to the state-of-the-art on the MPII, Human3.6M, and HumanEva-I datasets to prove the effectiveness of the proposed method.

4.4.2.5 Results on MPII

The PCKh0.5 metric score value has been reported in Table 4.5 for FSPE network on MPII dataset. This method is also compared with some of the state-of-the-art methods like Bulat et al. [149], Gkioxary et al. [201], Rafi et al. [185], Wei et al. [166], Chen et al. [186], Belagiannis et al. [158], Chu et al. [9], and Chou et al. [112]. We directly obtain their results from their published papers. Clearly, the FSPE method outperforms all these techniques under PCKh metric score.

To test the effectiveness of the method FSPE, we have used GT 2D pose instead of of the first stage of the given network for 3D reconstruction. A comparison chart is given in Figure 4.7 and Figure 4.8 compares the proposed FSPE (shown with blue color line) and GT 2D pose based 3D reconstruction (shown with orange color line) output over MPJPE and P-MPJPE evaluation criteria, which is 45.5 and 41.81 on average and 38.2 and 33.7 on average.

4.4.2.6 Results on Human3.6M

The validation of the effectiveness of the proposed 3D HPR method is discussed in this subsection.

This method gives 45.5, 38.2, and 45.8 reconstruction error over evaluation criteria-1, criteria-2, and criteria-3. The method outperforms the Pavllo et al. [142] (modeled on video input) by reducing the error by 1.3 mm over criteria-1 as shown in Table 4.6, Chen et al. [196] by 3.4 mm and 4.5 mm over Criteria-2 and Criteria-3 as shown in Table 4.7 and Table 4.8.

We also perform the evaluation separately for the second and third stage over MPJPE and P-MPJPE shown in Table 4.6 and Table 4.7. Therefore, it is verified that introducing the third stage benefited the 3D reconstruction by reducing the error. The third stage of the proposed network evaluated for different numbers of NSL value or frames. This proves that the NSL value 5 gives the best result rather than other numbers of frames as shown

in Figure 4.9. Next, we also observe that the method gives the best performance for evaluation Criteria-2.

4.4.2.7 Results on HumanEva-I

The 3D HPR method is evaluated for MPJPE error metric on HumanEva-I dataset, summarized in Table 4.9. The method is also compared with some of the state-of-the-art techniques shown in Table on Criteria 1(MPJPE). Few of the previous works have no results for boxing and jogging actions. So, they are reported as it is in the Table 4.7. Clearly, the proposed method shows the best performance on average MPJPE error value with 3.4 mm error reduction compared to other state-of-the-art methods.

4.4.2.8 Failure results

The main challenge that has been encountered with this method is that they function well for many poses but shows significant failure for certain types of poses like SittingDown and Sitting. Figure 4.10 shows the failure results occur in a few of the SittingDown pose class. In the future, this weak link of the proposed algorithm can be explored.

4.5 Conclusion

In this chapter, the proposed models attempted to solve the 3D HPE problem. The novelty of the methods lied in introduced frameworks. The models were compared against many

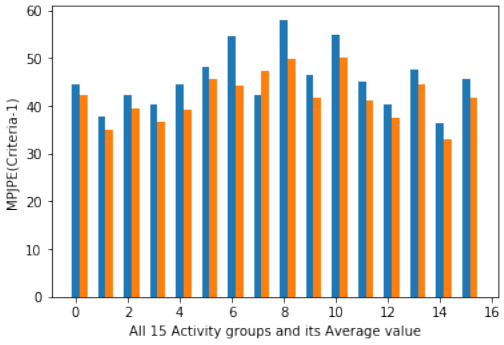

FIGURE 4.7: The MPJPE value on FSPE based (indicated by blue color) and 2D GT based (indicated by orange color) 3D HPR method having activity order mentioned in Table 4.2

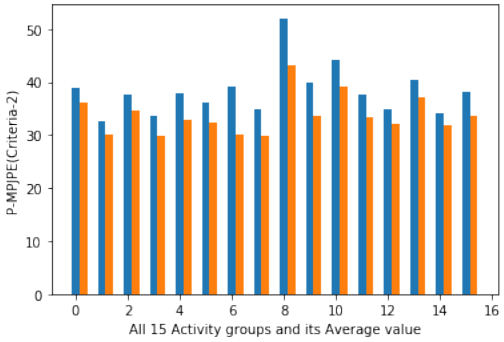

FIGURE 4.8: The P-MPJPE value on FSPE based (indicated by blue color) and 2D GT based (indicated by orange color) 3D HPR method having activity order mentioned in Table 4.2

Activity	Ours(NSL=5)	Ours(NSL=10)	Ours(NSL=20)
Discussion	44.4	40.15	40.31
Direction	37.7	35.53	36.64
Greet	42.1	39.73	40.56
Eat	40.3	42.38	45.41
Pose	44.4	40.02	37.88
Phone	48.2	51.41	53.04
Sitting	54.6	61.12	59.67
Purchase	42.1	38.7	41.21
SittingDown	58	62.18	68.5
Wait	46.4	43.48	46.13
Photo	54.9	52.75	54.15
Smoke	45.1	44.33	46.09
WalkTogether	40.4	41.67	43.34
WalkDog	47.6	45.45	49.88
Walk	36.4	34.53	35.31
Average	45.5	46.4	46.54

FIGURE 4.9: The MPJPE error value for different value of NSL =5, 10, and 20 with respective LSTM layers.

TABLE 4.6: The Human3.6M dataset has been utilized for 3D pose reconstruction errors(in millimeters) based on Evaluation Criteria 1.

Human 3.6M								
Method	**Discuss**	**Direct**	**Greet**	**Eating**	**Pose**	**Phone**	**Sitting**	**Purchase**
Criteria 1								
Methods uses Single Frame								
Zhang et al. [203]	59.90	52.83	61.95	61.58	57.03	85.47	81.29	58
Habibie et al. [204]	65.1	54.0	62.9	58.5	54.0	67.9	82.7	60.6
Moreno-Noguer et al. [205]	80.45	69.54	87.01	78.20	76.01	100.75	104.71	69.65
Pavlakos'17 et al. [11]	71.95	67.38	69.07	66.70	65.03	71.95	83.66	68.30
Pavlakos et al.[200]	54.4	48.5	52.0	54.4	49.9	59.4	65.8	52.9
Martinez et al. [114]	56.2	51.8	59.0	58.1	55.2	69.5	74.0	58.1
Liu et al.[206]	53.12	46.96	48.82	50.27	48.09	56.02	65.86	47.61
Yang et al.[13]	58.9	51.5	57.0	50.4	49.8	62.1	69.2	52.7
Methods uses more than one Frame								
Wang et al. [207]	59.96	50.03	56.55	54.66	52.74	65.65	85.85	54.81
Pavllo et al. [142]	46.7	45.2	45.6	43.3	44.6	48.1	57.3	44.3
Hossian & Little et al.[202]	50.7	48.4	55.2	57.2	53.0	63.1	66.1	51.7
Chen et al.[196]	44.2	41.1	45.9	44.9	41.6	46.5	73.2	54.8
Ours(Second-stage error)	44.5	40.7	44.2	42.5	45.09	49.22	57.5	44.66
Ours	44.4	37.7	42.1	40.3	44.4	48.2	54.6	42.1

	SittingDown	**wait**	**Photo**	**Smoke**	**WalkTogether**	**WalkDog**	**Walk**	**Average**
Methods uses Single Frame								
Zhang et al. [203]	98.29	63.29	81.32	68.27	53.79	65.83	49.42	66.55
Habibie et al. [204]	98.2	61.2	75.0	63.3	56.5	66.9	50.0	65.7
Moreno-Noguer et al. [205]	113.91	98.49	102.71	89.68	77.17	82.40	79.18	87.30
Pavlakos'17 et al. [11]	96.51	65.83	76.97	71.74	63.24	74.89	59.11	71.90
Pavlakos et al.[200]	71.1	52.9	65.3	56.6	47.8	60.9	44.7	56.2
Martinez et al. [114]	94.6	59.1	78.4	62.3	52.4	65.1	49.5	62.9
Liu et al. [206]	72.61	49.09	61.37	52.27	40.57	54.25	39.34	52.41
Yang et al.[13]	85.2	58.4	65.4	57.4	47.7	43.6	60.1	58.6
Methods uses more than one Frame								
Wang et al. [207]	117.98	59.55	79.63	62.48	48.52	65.21	41.48	63.67
Pavllo et al. [142]	65.8	44.0	55.1	47.1	33.9	49.0	44.0	46.8
Hossian & Little et al.[202]	80.9	57.3	72.6	59.3	49.6	62.4	46.6	58.3
Chen et al.[196]	46.2	42.1	39.3	48.7	38.5	35.8	46.6	46.3
Ours(Second-stage error)	60.30	46.65	55.55	47.96	41.6	48.2	38.04	47.11
Ours	58.0	46.4	54.9	45.1	40.4	47.6	36.4	45.5

state-of-the-art algorithms for MPII, Human3.6M, and HumanEva-I datasets. The performance evaluation was done based on MPJPE, P-MPJPE, and N-MPJPE error metrics. The proposed models were found better than state-of-the-art algorithms. The failure situations of the proposed algorithm were also discussed, which led to exploring strengthening other modules of the proposed algorithm. Further research will be focused on improving outcomes for these complex datasets. Multi-view data input-based algorithms have not been considered in this chapter.

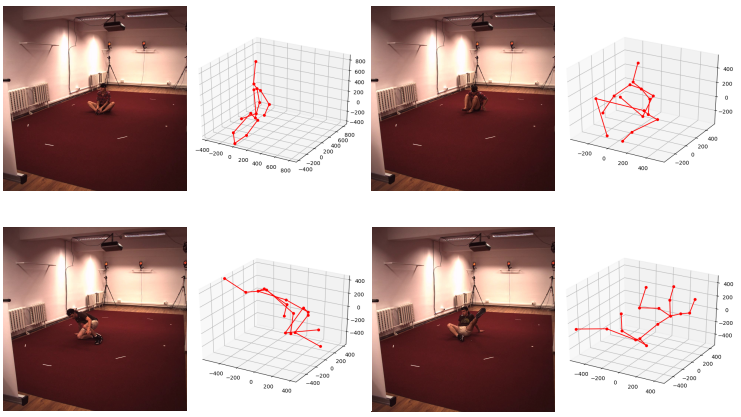

FIGURE 4.10: Few of the erraneous poses of the SittingDown Class

TABLE 4.7: The Human3.6M dataset has been utilized for 3D pose reconstruction errors(in millimeters) based on Evaluation Criteria 2.

Human 3.6M								
Method	**Discuss**	**Direct**	**Greet**	**Eating**	**Pose**	**Phone**	**Sitting**	**Purchase**
Criteria 2								
Methods uses Single Frame								
Pavlakos et al.[200]	54.4	48.5	52.0	54.4	49.9	59.4	65.8	52.9
Yang et al.[13]	30.9	26.9	39.9	36.3	28.8	43.9	36.9	29.4
Martinez et al. [114]	43.2	39.5	47.0	46.4	41.4	51.0	56.5	40.6
Methods uses more than one Frame								
Wang et al. [207]	59.71	48.54	56.12	56.12	57.31	67.68	78.26	55.57
Chen et al. [196]	39.3	36.9	41.2	40.5	38.0	42.0	67.5	51.2
Pavllo et al.[142]	36.1	34.1	37.2	34.4	34.4	36.4	45.0	33.6
Hossian & Little et al. [202]	37.9	36.9	40.3	42.8	37.7	46.8	48.9	36.5
Ours(Second-stage error)	37.90	33.7	39.4	34.0	38.33	37.7	39.6	35.87
Ours	38.8	32.6	37.5	33.6	37.8	36.0	39.2	34.9

	SittingDown	**wait**	**Photo**	**Smoke**	**WalkTogether**	**WalkDog**	**Walk**	**Average**
Methods uses Single Frame								
Pavlakos et al.[200]	56.8	39.6	42.5	42.6	36.5	43.9	32.1	41.8
Yang et al.[13]	58.4	30.5	47.4	41.5	32.2	29.5	42.5	37.7
Martinez et al. [114]	69.4	45.0	56.0	49.2	43.1	49.5	38.0	47.7
Methods uses more than one Frame								
Wang et al. [207]	115.85	61.29	71.47	69.99	51.42	62.22	44.63	63.74
Chen et al.[196]	42.1	37.5	34.9	42.5	34.2	30.6	40.2	41.6
Pavllo et al.[142]	52.5	33.8	42.2	37.4	27.3	37.8	25.6	36.5
Hossian & Little et al. [202]	52.6	39.6	46.7	45.6	38.5	43.5	35.2	42.0
Ours(Second-stage error)	53.4	40.0	45.9	37.80	35.10	41.02	35.7	38.98
Ours	52.0	39.8	44.1	37.5	34.9	40.3	34.1	38.2

TABLE 4.8: The Human3.6M dataset has been utilized for 3D pose reconstruction errors(in millimeters) based on Evaluation Criteria 3.

Human 3.6M								
Method	**Discuss**	**Direct**	**Greet**	**Eating**	**Pose**	**Phone**	**Sitting**	**Purchase**
Criteria 3								
Methods uses Single Frame								
Sun et al. [10]	50.5	52.4	57.8	45.0	46.1	49.8	96.3	57.1
Pavlakos et al. [200]	85.2	79.2	89.9	78.3	75.8	86.3	106.4	81.8
Methods uses more than one Frame								
Chen et al. [196]	48.0	45.9	50.8	48.6	46.1	48.9	57.4	64.73
Wang et al. [207]	45.62	38.69	48.92	54.77	47.49	77.3	54.65	47.17
Ours	46.8	33.9	43.3	41.3	44.6	48.8	57.6	44.0
	SittingDown	**Wait**	**Photo**	**Smoke**	**WalkTogether**	**WalkDog**	**Walk**	**Average**
Methods uses Single Frame								
Sun et al. [10]	47.4	52.1	50.3	56.4	48.7	45.7	53.7	53.6
Pavlakos et al. [200]	137.6	92.3	87.9	86.2	77.5	72.9	82.3	88.6
Methods uses more than one Frame								
Chen et al. [196]	49.4	47.2	45.1	54.2	42.9	39.9	49.9	50.3
Wang et al. [207]	94.30	49.29	78.85	56.84	38.96	58.71	33.07	54.14
Ours	48.8	57.6	44.0	55.1	44.3	46.8	38.8	45.8

TABLE 4.9: The HumanEva-I dataset has been utilized for 3D pose reconstruction errors(in millimeters) based on Evaluation Criteria 1.

	Jogging				**Walking**				**Boxing**			
Method	S1	S2	S3	Avg.	S1	S2	S3	Avg.	S1	S2	S3	Avg.
Methods uses Single Frame												
Pavlakos et al. [200]	23.5	15.4	14.5	17.8	18.8	12.7	29.2	20.2	-	-	-	-
Pavlakos'17 et al. [11]	28.9	21.9	23.8	24.9	22.3	19.5	29.7	23.8	-	-	-	-
Moreno-Noguer et al. [205]	39.7	20.0	21.0	26.9	19.7	13.0	24.9	19.21	-	-	-	-
Martinez et al. [114]	26.9	18.2	18.6	21.2	19.7	17.4	46.8	27,9	-	-	-	-
Yasin et al. [12]	46.6	41.4	35.4	38.9	35.8	32.4	41.6	36.6	-	-	-	-
Tekin et al. [208]	-	-	-	-	37.5	25.1	49.2	37.3	50.5	61.7	57.5	56.6
Methods uses more than one frame												
Wang et al. [207]	27.9	19.5	20.9	22.8	17.2	13.4	20.5	17.0	29.7	44.0	47.2	40.3
Lin et al. [209]	41.0	29.7	29.1	33.2	26.5	20.7	38.0	28.4	39.4	57.8	61.2	52.8
Ours	22.7	14.5	13.4	16.8	16.7	11.9	19.0	15.86	25.5	40.0	45.4	36.96