Chapter 5

DEEP NETWORK FOR 3D HUMAN POSE ESTIMATION USING MULTI-VIEW DATA

5.1 Introduction

In numerous robotics and computer vision (CV) applications, for instance such as augmented and virtual reality, self-governing autonomatic driving, understanding of spatial data of the subjects in a frame or scene is important. Poor knowledge of depth and spatial data arrangement can considerably restrict the algorithms performance. In this paper, the authors examine an specific instance of depth and spatial understanding: 3D HPR from images and videos.

To estimate human pose as 3D from 2D data representations is very well known challenging and active research area among the CV and graphics community. An awareness of the human body poses information and the limb connection knowledge is essential for higher-level tasks of CV like human activity or action identification [210] [211], virtual and augmented reality, sports analysis.

Over the years, several different procedures have been utilized to handle the task of 3D HPR from images and videos. To perform 3D reconstruction, mapping has been performed from an image to 3D representation, this approach needs to be invariant to the many number of factors, including image imperfections, texture of cloth, color of skin, lighting, and background scenes. Prior to coming deep neural systems, many of the procedures tend to use hand-engineered descriptors, like SIFT descriptors [212], shape context [213], edge direction histograms [99], silhouettes [104] or to learn a model that can estimate the 3D poses from frames.

Due to the recent progress of the Deep Learning approach in the area of CV, numerous systems have attempted to exploit the powerful discriminative ability of deep networks to estimate 3D poses. Most of the recent techniques based on deep networks utilized a single-view image for 3D reconstruction [11] [114], due to the immense flexibility of the depth knowledge extracted from a single-view frame. Although these techniques give acceptable performance using a single-view image, it is an ill-posed problem due to some factors like clothing, variation in human appearance, or self-occlusion and still require to improve the reconstruction accuracy.

Over years, two principal kinds of methods have been submitted: Discriminative-based ones that, directly, estimate the 3D poses from RGB image [11] [214] [215] [105] and generative-based ones that, firstly, estimate the 2D pose to reconstruct the 3D pose later [114] [113] [216]. Many of the recent systems have tried to straightly estimate the 3D poses from RGB images by training the architecture end-to-end [11] [214] [215] [105]. Various CV systems oriented on deep neural models presently give better results compared to the traditional approaches on objectives like 2D pose estimation [111] [166], object classification and localization [217] [218]. However, deep neural models demand a large number of input data to function perform well. However, contrary object or frame classification or 2D HPE systems that have abundant data, there is a lack of ground truth (GT) 3D human pose data. This causes the task of inferring the 3D poses straight from the image extremely challenging. Even though a few end-to-end algorithms for 3D HPR have provided remarkably better results compared with many of the older approaches, the primary source of error in such systems are not well studied and understood. The only factors that make the system erroneous are image imperfection, skin color, lighting, clothing texture, and shape.

Many recent techniques exploit 2D pose information for 3D reconstruction [114] [113] [13]. While dividing the problem might reduce the difficulty of the task, it is inherently ambiguous since multiple 3D poses can be mapped to the same 2D pose. On the other hand, this type of approach provides the invariance to the above-mentioned factors. Therefore, we can see that both techniques have their pros and cons. However, if both the approaches are integrated [219] [109], the direct method behaves reliable and has more

merits.

Above discussion motivates us, in this paper, we attempt to utilize the best of both the systems in a multi-view scenario as shown in Figure 1, to build the system additionally accurate towards the reconstruction accuracy.

In summary, The above-discussed factors motivate us to propose a two-stage multi-view deep network for 3D HPR. The first stage predicts the 2D pose heatmaps using the proposed enhanced stacked-hourglass technique from all the multi-view images. The second stage utilizes both the generated heatmaps and the original image with the proposed multi-view 3D reconstruction deep network. This network utilizes both the heatmaps and the original images in the given Heatmap Flow and Image Flow with early and late fusion strategies to reconstruct the 3D pose. The Heatmap Flow of this stage utilizes the generated 2D joint heatmaps to reconstruct a 3D pose. The Image Flow used to generate features that are the complement of those computed by the Heatmap Flow and used in conjunction with them to reconstruct the 3D pose. The novelty of our work lies in exploiting both the image and 2D heatmap information in the multi-view scenario to reconstruct the 3D pose.

Ultimately, the proposed method supports the system to yet utilize image information while reconstructing 3D poses from 2D poses. Through the experimental demonstration, we analyze that the calculated features in both the flow are highly decorrelated that's why they encode complementary data cues.

The contributions to this chapter are:

- Introduced a two-stage multi-view 3D HPR technique using deep learning, which gives high accuracy as compared to other recent state-of-the-art techniques.

- Proposed an enhanced stack-hourglass approach (ESHA) approach for 2D HPE that uses the backbone of the stackhourglass technique. It gives a more accurate prediction in terms of the PCK metric compared to stackhourglass.

- Proposed a multi-view spatial 3D reconstruction deep network that utilizes both the direct image and the 2D pose information in a multi-view scenario to reconstruct the 3D pose.

5.2 Related Work

The multi-view based 3D HPR has largly utilized nowadays in the field of CV and have more chances to obtain a huge amount of work in the literature. For this cause, we define our bibliographical outline solely to practices that are likewise correlated to the proposed strategy, i.e., methods use the deep neural network and handling a multiview scenario.

Hong et al. [106] introduced a 3D HPR approach that utilizes the 2D silhouettes for reconstruction. Here the authors also introduce a multimodal deep autoencoder (MDA) and also backpropagation NN (BP-NN). Then implements a fusion over the features and later on perform the mapping from silhouettes to 3D pose. Hang et al. [220] also proposed a method that utilizes the local similarity preservation concept to collect the alike silhouettes to alleviate the uncertainty of its sparse codes. The use of sparse code in a multi-view

scenario here enhances its performance. Hong et al. [221] introduced to extract feature by using deep learning techniques over a denoising autoencoder, that contains manifold regularization along with hypergraph laplacian. In this paper Hong et al. [221] mainly worked to extract automatic features for silhouettes.

Trumble et al. [6] introduce an approach for combining multi-view video data along with inertial measurement unit sensor data. They suggest a 3D DCNN that uses a volumetric probabilistic based visual hull to get a pose embedding from the videos. The suggested a DCNN model having a line of 3D C layer followed by the max-pooling layer succeeded with the set of FC layers. Then, the model fuses this outcome to the inertial sensor data and supplies to the LSTM-based deep network.

Pavlakos et al. [50] give a geometry-driven procedure for the task of pose prediction. They introduce a cascaded arrangement that begins by a generic CNN to estimate the probability maps of the 2D pose. Later, the authors introduce an automated approach that uses probability maps to accumulate reliable 3D pose annotations. They utilize the human body 3D structure and calibration constraints to consolidate the 2D CNN estimations in every view to reconstructing the 3D pose.

Ershadi-Nasab et al. [222] recommend a system that utilizes data from multiview frames for the reconstruction of 3D human pose. They use the FC pairwise random field comprising of two pairwise names. The first one uses articulated body configuration to encode the spatial dependence of the joints. Another one is founded on the final product of the deep

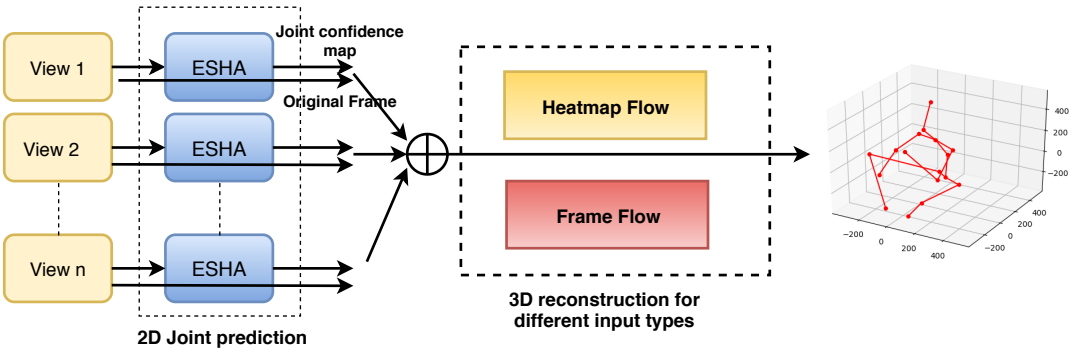

FIGURE 5.1: The framework of the proposed method.

2D part estimator by Insafutdinov et al. [184]. At last, the method presents the outcome by applying the Loopy Belief Propagation approach.

Nunez et al. [216] introduced a deep neural network for 3D human pose reconstruction based on the multi-view images. The authors first estimate the 2D joint coordinate for each view. Then theses 2D poses have been used to reconstruct 3D pose with the introduced least-square optimization approach. Lastly, LSTM has been used for temporal refinement of the 3D pose.

Instead of using both Image and 2D pose information for 3D HPR, the discussed methods only utilize either of the data for reconstruction.

The remaining paper is arranged as follows: Section 3, discusses the proposed method for 3D HPR. Section 4 describes the experimental evaluation of the proposed method. At the end Section 5 convey the conclusion of the whole work.

5.3 Proposed Method

In this paper, the task is to reconstruct the 3D human pose using the multi-view images. The proposed method has arranged in three consecutive parts as shown in Figure 1. The first part of the method estimates the 2D joint heatmaps for each view separately using the proposed Enhanced stack-hourglass approach (ESHA). The second part uses both the original image and these 2D heatmaps from all the obtainable views to reconstruct a 3D pose. Finally, in the third part, an LSTM deep module has been utilized to improve the reconstruction accuracy. The input is the set of multi-view images $\{b_{im}\}_{im=1}^{n}$ and the output is set of 3D joint coordinates which is represented as $\left\{X^k\right\}_{k=1}^{K}$. The detailed description of each part is given in the below subsections.

5.3.1 Enhanced stack-hourglass approach (ESHA)

In this, we propose a CNN based 2D Human Pose Estimation (HPE) approach from an frame [223]. The model has two parts of feature extraction (FE) and its refinement. The method first performs the FE over the image that extracts the basic features. The stem part of inception-v4 [167] is utilized here for FE. The inception neural deep model is very tunable, so that authors can efficiently formulate modifications on the filter count level at various layers, so it does not effect the fully trained model peformance efficiency and taking out important information that build the estimated output accurate benefitted with low compute cost load. We observe that this CNN module for FE has rich quality data information to prepare an precise 2D prediction.

The derived feature information from the FE part have been passes as a input to the feature refinement (FR) module which has been based on the stackhourglass technique [111]. HPE is highly demanding than of image classification. Later on, after the introduction of the stack-hourglass approach, varied approaches in published works utilize the multi-level network for HPE. So, to take benefit from this observation we employ four stages of the hourglass for FR in proposed model.

We examine that the stackhourglass procedure has been inadequate for HPE for the reason of its model's layout choice, they make use of recurring up-down sampling, that posses huge risk of information loss. The authors try work to overcome this flaw by giving a Multi-stage Cascaded Feature Connection (MSCFC) procedure, across each stage of the FR having four stages. This MSCFC strategy move around the diverse scale data, which extenuate the training hurdle and reinforces the data flow. For all possible individual scale, two distinct data flows have been suggested from down sampling and up sampling parts of the preceding stage to the down sampling part of the present existing stage.

The 1×1 convolve operation has been attached on all possible flow. The given architecture model gives joint heatmaps as output. All output of 2D HPE prediction has been illustrated as the K heatmaps, which denotes the total number of body joints. The $b_{im} \epsilon R^{W \times H \times 3}$ is the data as RGB frame for im^{th} view. The $hm_k^{im} \epsilon R^{W_{hm} \times H_{hm} \times L}$ (k= 1,2,....,K) is the K^{th} joint heatmap from im^{th} view. This whole technique for view i is mapped like:

$$f(y^{im}) = ((hm_1^{im}, ..., hm_k^{im})) \tag{5.1}$$

The loss oriented with the heatmaps for in-between supervision is as:

$$Loss_{2D}^{im} = 1 \div k(\sum_{k=1}^{K} \left\| hm_k^{im} - \hat{hm_k^{im}} \right\|) \tag{5.2}$$

$\hat{hm_k^{im}}$ is the 2D GT through a Gaussian kernel with mean equal to the GT and variance one.

Algorithm 4: ESHA for 2D pose prediction

Input: Frame
Output: 2D Pose
Procedure
Feature Extraction:
$Input \leftarrow frame : x$
$x' \leftarrow Inception_{stem}(x)$
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**

5.4 Multi-view spatial 3D reconstruction approach

The goal of this part is to perform the reconstruction for 3D HPR by using both the image and 2D joint heatmaps in a multi-view scenario. We designed a two-flow deep network for

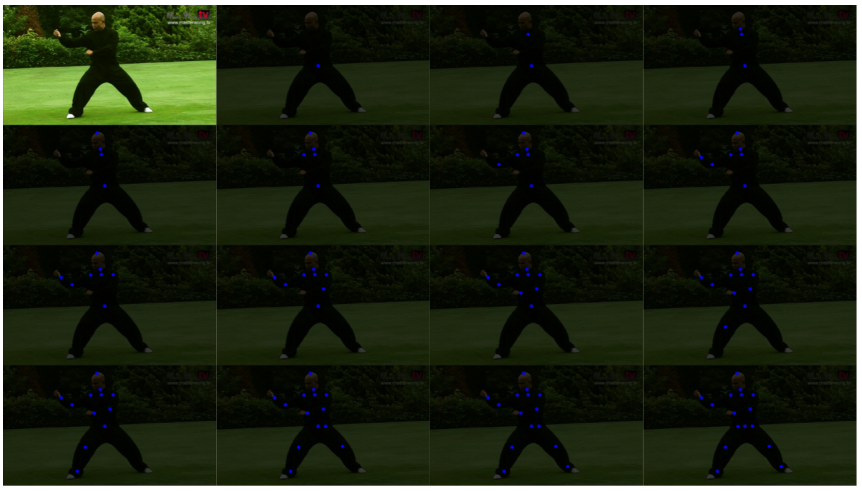

FIGURE 5.2: The heatmaps for different joints.

this aim as shown in Figure 5.2. The input of this part is the multi-view images and its 2D heatmaps $\sum_{im=1}^{n} b_{im}, \sum_{im=1}^{n}(hm_1^{im}, ..., hm_k^{im})$ and the output is is 3D joint coordinates $\left\{X^k\right\}_{k=1}^{K}$. The heatmap flow part first takes the 2D joint heatmaps from all the views to compute the feature maps. The image flow uses all the multi-view images to extracts the additional features directly. Then all these features have been fused to reconstruct 3D human pose. Here we perform early and late fusion concepts over these flows.

Figure 5.2 and 5.3 shows the basic building block of the proposed deep network, utilizes the convolutional deep network having BN [224], D [225], RELU [226] and residual module [173].

Convolutional(C)-ReLU(R) layers The convolutional neural networks, is used because this network easily learns the translation-invariant filters which are used with the image and joint-location heatmaps. RELU is used to incorporate the non-linearity in the systems.

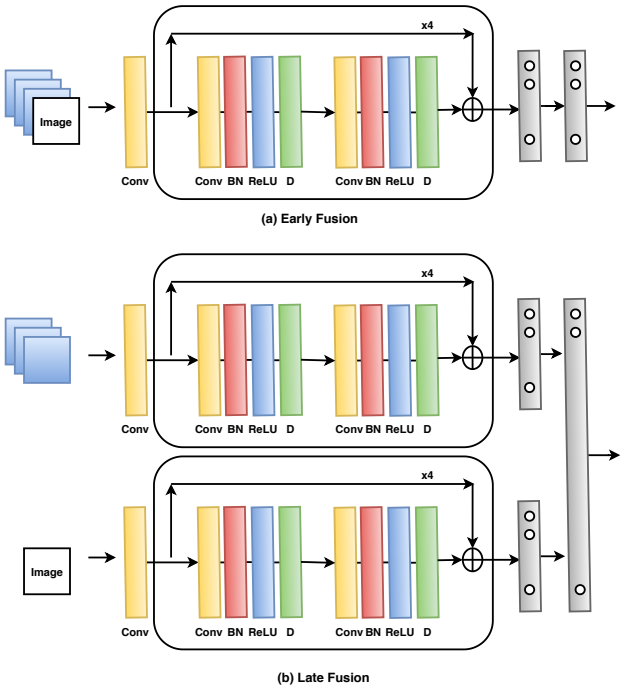

FIGURE 5.3: Early and Late fusion network over Heatmap and Frame Flow

Mathematically, the convolutional operation is the measure of the overlap amount between two functions. It can be expressed as follows: For 1-D:

$$z(i) = (v * w)(i) = \int w(n)v(i-n)dn \tag{5.3}$$

In descrete terms this can be expressed as:

$$z(i) = (v * w)(i) = \sum_{n} w(n)v(i-n) \tag{5.4}$$

For 2D-spatial data:

$$z(i) = (v * w)(i,j) = \sum_{m}\sum_{n} w(m,n)v(i-m,j-n) \tag{5.5}$$

$*$ is the convolution operation. s denotes the input, k is the kernel and r is the output. Here m and n is the count of rows and columns of an image. The ReLU operation is expressed as:

$$\theta(s) = max(0, s) \tag{5.6}$$

ReLu gives a positive output s if s>0and 0 otherwise. ReLU is nonlinear and so layers can be effective in the network. Further, given how a network is typically randomly initialized with weights, ReLU results in far fewer activations to process and this makes it less computationally costly.

Batch Normalization(BN) and Dropout(D) Before these two concepts system do not give a satisfactory result. After applying BN and D, it improves the performance of the 3D HPR while having slightly more time in train and test.

BN is a way to give any layer in a CNN with data that are unit variance/zero mean. The BN algorithm follows the given below equations(Input: Values of x over a mini-batch: β = $x_{1,..,m}$, parameters: β, γ; Output: $y_i = BN_{\gamma,\beta}(x_i)$):

$$\mu_\beta \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \tag{5.7}$$

$$\sigma_\beta^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_\beta)^2 \tag{5.8}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}} \tag{5.9}$$

$$y_i \leftarrow \gamma\hat{x}_i + \beta = BN_{\gamma,\beta}(x, i) \tag{5.10}$$

A dropout factor specifies the probability at which outputs of the layer are dropped out, or inversely, the probability at which outputs of the layer are retained.

Residual Module [173] motivates to use the residual connection because this makes easier the training of the very deep CNN. Here it facilitates the system to reduce the error by four percentage.

Image/2D joint heatmaps and its fusion Image Flow oriented feature maps and Heatmap Flow based feature maps are expressed as $\{I_j\}_{j=0/L}$ and $\{H_j\}_{j=0/L}$. Here j represents the layer. These feature maps H_j and I_j for j=0/L coincide in hieght and width but having distinct channel count. The input for $(j+1)^{th}$ layer is same as the output of the j^{th} layer.

Let $\{U_j\}_{j=0/L}$ is the feature maps of the fusion system. Here U_j denotes the layer j output. For j=L case, the layer j input is the linear combination of U_j with H_j and I_j is expressed as:

$$(1 - w_j) \cdot fusion(I_j, H_j) + w_j \cdot U_j \tag{5.11}$$

here fusion Fusion($\cdot$, $\cdot$) is the simple concatenation operation along the channel axis over the feature maps and w_j is the weight $w \in [0, 1]^L$ that controls the fusion. To make this fusion possible, the size of U_j should be same as the size of H_l and I_l and the channel count is equal to the summation of both H_l and I_l channel count. For specific case, $U_0 = fusion(I_0, H_0)$, and $U_{L+1} \in R^{3k}$ is the representation of output of the system. The different setting of w leads to different fusion strategies. We discuss below the two used fusion cases:

The denotion of input for early and late fusion is given as:

$$\sum_{i=1}^{I}\sum_{k=1}^{K}(H_k^i, I^i) \tag{5.12}$$

$$\sum_{i=1}^{I}\sum_{k=1}^{K} H_k^i, \sum_{i=1}^{I}\sum_{k=1}^{K} I^i \tag{5.13}$$

Here H_k^i denotes the generated heatmaps for i^{th} view, I^i denotes the image for i^{th} view.

Early Fusion The early fusion strategy utilizes the single flow, where the image and heatmaps are fused at initial input level to perform the computation. Here U_0 is concat(I_0,H_0).

Late Fusion Here the fusion of both the flow occurs at the last layer. The fused layer is the linear concatenation of the both image and heatmap flow FC outputs.

The multi-view mapping system with the proposed flows has been mapped as:

$$(\hat{q}) = Multiview_{3D}(set((H_1^1, ..., H_1^I), set(H_k^1, ..., H_k^I)), \\ set(I_1, ..., I_I)) \tag{5.14}$$

The loss function for 3D has been defined as:

$$L_{3D}^v = 1 \setminus j \sum_{j=1}^{J} \left\| q_j^v - \hat{q}_j^v \right\| \tag{5.15}$$

Here q_j^v is the estimated output and $\hat{q}_j^v$ is the GT for that.

Input: Multi-view Images and its 2D Heatmaps $\sum_{im=1}^{n} b_{im}, \sum_{im=1}^{n}(hm_1^{im}, \ldots, hm_k^{im})$

Output: 3D Pose: $\left\{X^k\right\}_{k=1}^{K}$

Early Fusion

Input: $x : (((k+n)+3) \times W \times H)size$,

ANOTHER THING A

x ← C(x)

for $i \leftarrow 0$ **to** 0 ***By*** 4 **do**

  x ← C(x)

  x ← BN(x)

  x ← R(x)

  x ← D(x)

  x ← C(x)

  x ← BN(x)

  x ← R(x)

  x ← D(x)

end

x ← FC(x)

x ← FC(x)

return *x*

Late Fusion

Input: $x : (((n)+3) \times W \times H)size$,

$y : (((n)+k) \times W \times H)\ A$

Step 1 to 12 applied over both x and y

independently x← C(x)

for $i \leftarrow 0$ **to** 0 ***By*** 4 **do**

  x← C(x)

  x← BN(x)

  x← R(x)

  x← D(x)

  x← C(x)

  x← BN(x)

  x← R(x)

  x← D(x)

end

x ← FC(x)

x1 = Concatenate(x,y)

x ← FC(x1)

return *x*

3 5.5 Experiments

In this segment of the paper, we discuss and present the computational experimental operations implemented to evaluate the efficiency of the proposed method and also examine it to techniques recognized in the literature.

5.5.1 Dataset and Evaluation protocol for ESHA

MPII Dataset. MPII [127] is the single person HPE benchmark dataset. The dataset contains the group of images from the videos taken from Youtube that include every day human actions. It comprises 25,000 images together with 40,000 annotations. The 30,000 from them have been employed for training and the rest of 10,000 is for testing.

Evaluation Metric(PCKh) First, we talk about the PCK metric. Here the estimated prediction of the joint is handled correctly if the distance of the GT and prediction output rests inside the inclined threshold. In this work, the threshold is taken as 0.5 and it is interpreted as PCK@0.5.

A different modification of PCK is termed as PCKh, where the estimated prediction of the keypoints is considered as right if the keypoint gap of the distance between calculated and GT is under than the α part of head segment length (PCKh). The general notation of this metric measure is as PCKhα.

TABLE 5.1: The ESHA algorithm result and its comparability with distinct state-of-the-art procedures for the basis evaluation metric PCKh with @0.5.

Algorithms	Elbow	Knee	Shoulder	Head	Ankle	Hip	Wrist	Total
Gkioxary et al. [201]	86.7	81.4	93.1	96.2	74.1	85.2	82.1	86.1
Wei et al. [166]	88.7	83.4	95.0	97.8	78.0	88.4	84.0	88.5
Rafi et al. [185]	86.4	80.6	93.9	97.2	73.4	86.8	81.3	86.3
Belagiannis et al. [158]	88.2	82.6	95.0	97.7	78.4	87.9	83.0	88.1
Chen et al. [186]	92.5	89.6	96.5	98.5	86.0	90.2	88.5	91.9
Chou et al. [112]	92.2	89.1	96.8	98.2	84.9	91.3	88.0	91.8
Bulat et al.[149]	89.4	97.9	85.3	95.1	89.9	81.7	85.7	89.7
Chu et al. [9]	89.9	85.7	95.1	97.9	81.7	89.4	85.3	91.5
Ours	94.1	89.9	98.5	98.9	85.2	92.6	86.0	92.2

5.5.2 Datasets and Evaluation protocol for 3D HPR

The experimental evaluation for the 3D based HPR system has been executed on the publicly accessible Human3.6M 3D dataset.

Evaluation Metric for 3D HPR

- **Metric 1(MPJPE Error Metric):** MPJPE is stated as the mean of the per joint position error(PJPE). The PJPE is calculated using the Euclidean distance between the reconstructed 3D pose and the GT 3D pose. The mean of the calculated distance is called here as an MPJPE metric. Let the number of skeleton joints are N_{sk}, frame and skeleton is represented by fr and sk.

$$E_{MPJPE}(fr, sk) = \frac{1}{N_{sk}} \sum_{i=1}^{N_{sk}} \left\| u^{fr}_{fr,sk}(i) - u^{fr}_{gr,sk}(i) \right\|_2 \tag{5.16}$$

- **Metric 2(PA-MPJPE Error Metric):** Another type of traditional way that works to evaluate their processes is to do the alignment of an estimated 3D human pose

with the GT doing a similarity transformation (Procrustes analysis).

$$E_{PA-MPJPE}(fr, sk) = \frac{1}{N_{sk}} \sum_{i=1}^{N_{sk}} \left\| u - aligned^{fr}_{fr,sk}(i) - u^{fr}_{gr,sk}(i) \right\|_2 \quad (5.17)$$

where u-aligned is the Procrustes alignment of the predicted joints to the correspondent GT of the joints.

- **Metric 3 (N-MPJPE Error Metric):** In this metric before calculating MPJPE, aligment is done using least-square method on GT with respect of scale.

Human3.6M Dataset. This dataset is the largest one utilized for multi-view 3D HPR. The dataset contains a total of 3.6 million frames taken by 15 actions performed by 7 different actors in the constraint of the laboratory environment. All the actions were recorded using a four-view camera organization having both extrinsic and intrinsic parameters.

Evaluation-protocol#1 Here the respective five subjects (S1, S5, S6, S7, and S8) have been used for training and (S9 and S11) used for testing. This evaluation protocol has been largely used by many recent techniques [11] [114].

5.5.3 Implementation Details

In this part of the chapter, we introduce and explain the computational examinations conducted to examine the effectiveness of the proposed two-stage deep network and also compare it to systems recognized in literature history.

The training and testing of the proposed two-stage deep network have been conducted on the computer with a setup structure of 12GB NVIDIA Tesla K40c. For earliest stage, the data input image was resized and cropped to 256x256 pixels considering the human at the middle or center of the frame for the MPII 2D dataset. To perform data augmentation we apply random rotation (+- 40 degree), random data augmentation with translation (associated with +-2% of the stated image data size), and then perform scaling from 0.7 to 1.3 on dataset MPII. For the demonstration, we use 1e-3 of base learning rate. This drop out by a part of 0.4 whenever the loss for the validation data set saturates. The authors utilized the validation split as same as [8].

For the Human3.6M dataset, We have trained the proposed model for 110 epochs, with each epoch going through the entire human 3.6 million records. We used Adam [227] optimizer to train the model by a learning rate of 1e-6, which drops off quickly per iteration.

5.5.4 Results and Discussions

This section explains the results achieved by the introduced approach. We also give a comparison table with the state-of-the-art literature methods for the MPII 2D dataset, and Human3.6M 3D dataset to demonstrate the effectiveness of the introduced technique. This section also discusses the weakness of the given algorithm. The visualization of the 3D output is shown in Fig. 5.4.

We have performed the experiments to check the effectiveness of the proposed method on different resolution input and illumination shown in Fig. 6.6 and Fig. 5.5. We have also evaluate the result on image having more than one person as shown in Fig. 5.7. These experiments prove that the proposed method largly invariant to varrying resolution, illumination and number of person in the data. The computational load of the proposed method in terms of time is 0.331 sec for 2D joint locations estimation and 0.002 for 3D HPR.

5.5.4.1 Results and comparison with state-of-the-art

MPII Dataset. In this subsection, we do the comparison of the proposed ESHA network with the baseline stack-hourglass method as shown in fig in terms of the PCKh evaluation metric. The method also compared with some other state of the art methods and shows its significance superiority over them. This compared state-of-the-art procedures are Gkioxary et al. [201], Wei et al. [166], Rafi et al. [185], Belagiannis et al. [158], Chen et al. [186], Chou et al. [112], Bulat et al.[149] and, Chu et al. [9]. We straightly take their score value from their published articles. The ESHA approach beats all these systems following the PCKh metric value score.

Human3.6M Dataset. The error output of the proposed deep network and the state-of-the-art techniques are given in Table 5.2, 5.3 and 5.4. The table 5.2 compares the proposed method for the MPJPE reconstruction error for four sets of views with Zhang et al. [203], Wang et al. [207], Chen et al.[196], Habibie et al. [204], Pavllo et al. [142], Lee

et al.[199], Pavlakos et al.[200], Hossian & Little et al.[202] and, Yang et al.[13]. The results of other state of the art methods are taken directly from the respective articles. The result under the protocol1 and evaluation metric MPJPE clearly shows that the method outperforms all these competing techniques. For a fair comparison, we only consider those state-of-the-art methods that work for the same evaluation protocol and metric.

Similarly, the Table 5.3 shows the significant superiority of the method for the PA-MPJPE metric with Wang et al. [207], Pavllo et al.[142], Chen et al. [196], Pavlakos et al.[200], Hossian & Little et al. [202] and, Yang et al.[13] techniques. The Table 5.4 shows the result of the introduced approach on N-MPJPE metric which is named as Criteria 3 and its compararison with Wang et al. [207], Chen et al.[196], and Pavlakos et al.[200].

In summary, we believe that the presented method indicates clear evidence that the advantage of fusing the heatmap and image information in the multi-view scenario, as done by us.

5.5.4.2 False Results

The major difficulty that has been found with proposed approach is that the model function satisfactory for majority of the poses but presents significant failure output for the few kinds of poses such as SittingDown (SittingD) and Sitting as shown in Fig. 5.5. For the future, we work on this flawed part of the introduced procedure.

5.6 Conclusion

In this chapter, we introduced a multi-view deep network to reconstruct a 3D pose. We presented this solution by exploiting both 2D joint heatmaps and its original image information in a multi-view scenario to reconstruct the 3D pose. In part of this method, we predict 2D joint heatmaps using the proposed ESHA, which is based on the stack-hourglass method. The resulted heatmaps have been fused with the original image in multi-view senario for 3D reconstruction. The novelty of the method lies in the introduction of ESHA method for 2D pose and utilization of both the data information in multi-view setup. Experimental outcomes prove that the proposed design choice of the deep network can significantly enhance the performance on the Human3.6M dataset in terms of MPJPE and PA-MPJPE error metric. Through these experimental outcomes, we also proved that utilizing a fusion of the 2D joint heatmap and image information in the multi-view scenario enhances the 3D HPR functioning in the form of the reconstruction error. In the future job, we intend to examine this type of fusion to reconstruct 3D HPR in-the-wild videos.

5.7 conclusion

In this chapter, we introduced a multi-view deep network to reconstruct a 3D pose. We presented this solution by exploiting both 2D joint heatmaps and its original image information in a multi-view scenario to reconstruct the 3D pose. In part of this method,

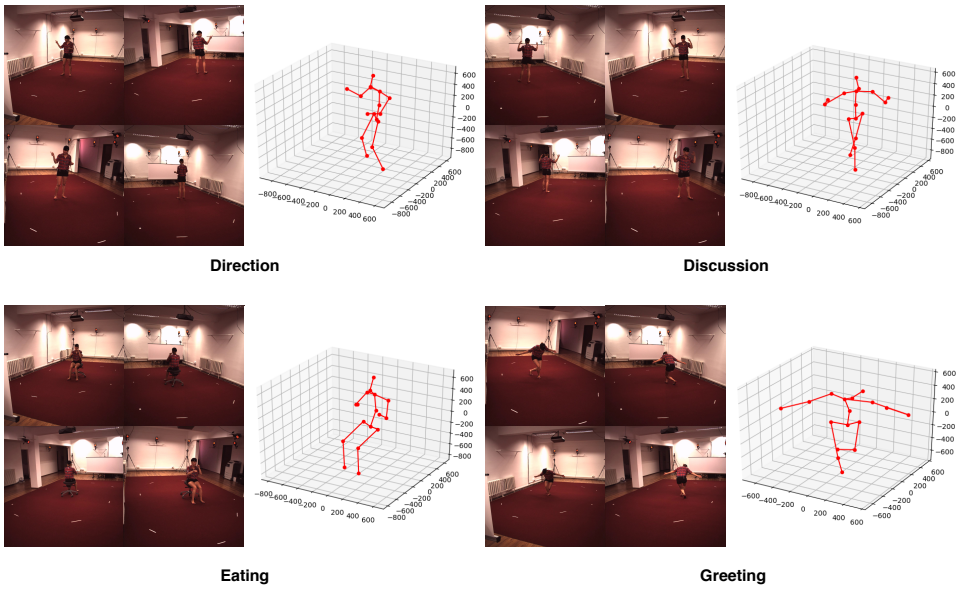

FIGURE 5.4: Multi-view images and its 3D pose.

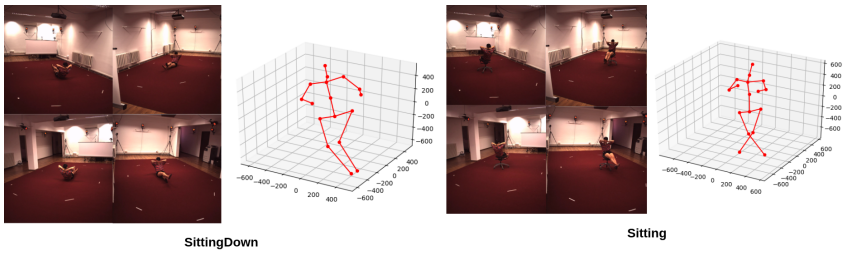

FIGURE 5.5: Failure Cases

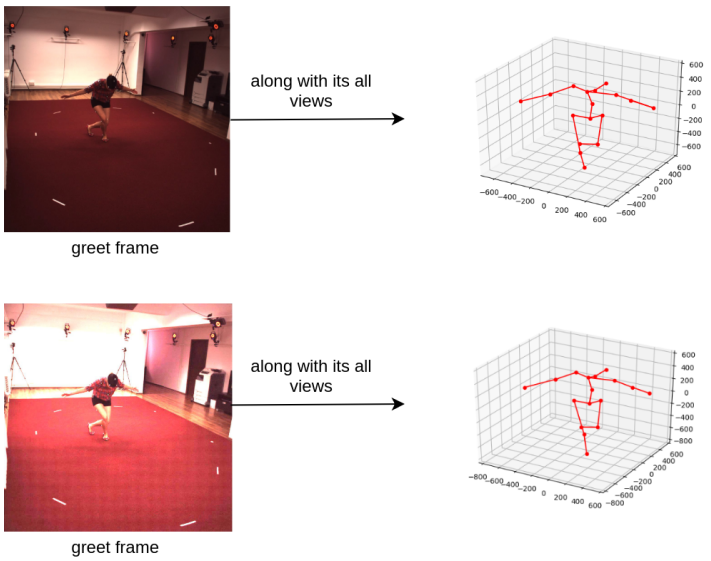

FIGURE 5.6: Illumination change output

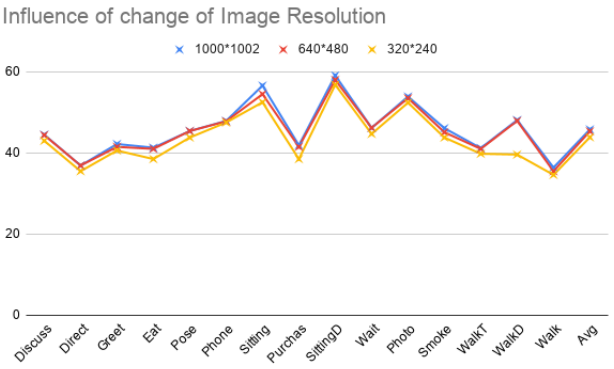

FIGURE 5.7: The influence of change of image resolution.

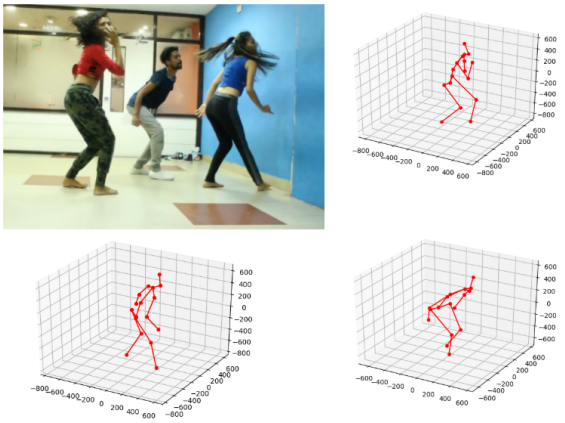

FIGURE 5.8: Early and Late fusion network over Heatmap and Frame Flow

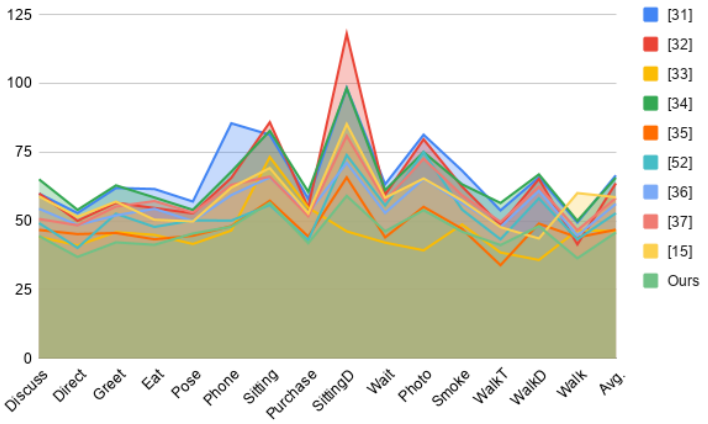

FIGURE 5.9: Comparision chart for reconstruction error

TABLE 5.2: The Evaluation Criteria 1 over Human3.6M public data set has been employed to calculate the 3D HPR errors (in the form of millimeters).

Human 3.6M								
Algorithm	**Greet**	**Direct**	**Discuss**	**Phone**	**Pose**	**Eat**	**SittingD**	**Purchase**
Criteria 1								
Zhang et al. [203]	61.95	52.83	59.90	85.47	57.03	61.58	98.29	58
Wang et al. [207]	56.55	50.03	59.96	65.65	52.74	54.66	117.98	54.81
Chen et al.[196]	45.9	41.1	44.2	46.5	41.6	44.9	46.2	54.8
Habibie et al. [204]	62.9	54.0	65.1	67.9	54.0	58.5	98.2	60.6
Pavllo et al. [142]	45.6	45.2	46.7	48.1	44.6	43.3	65.8	44.3
Lee et al.[199]	52.6	40.2	49.2	50.1	50.2	47.8	73.9	43.0
Pavlakos et al.[200]	52.0	48.5	54.4	59.4	49.9	54.4	71.1	52.9
Hossian & Little et al.[202]	55.2	48.4	50.7	63.1	53.0	57.2	80.9	51.7
Yang et al.[13]	57.0	51.5	58.9	62.1	49.8	50.4	85.2	52.7
Ours	44.5	36.9	42.2	41.3	45.4	47.9	56.6	41.9

	Sitting	**Smoke**	**Photo**	**Wait**	**Walk**	**WalkDog**	**WalkT**	**Average**
Zhang et al. [203]	81.29	68.27	81.32	63.29	49.42	65.83	53.79	66.55
Wang et al. [207]	85.85	62.48	79.63	59.55	41.48	65.21	48.52	63.67
Chen et al.[196]	73.2	48.7	39.3	42.1	46.6	35.8	38.5	46.3
Habibie et al. [204]	82.7	63.3	75.0	61.2	50.0	66.9	56.5	65.7
Pavllo et al. [142]	57.3	47.1	55.1	44.0	44.0	49.0	33.9	46.8
Lee et al.[199]	55.8	54.1	75.0	55.6	43.3	58.2	43.3	52.8
Pavlakos et al.[200]	65.8	56.6	65.3	52.9	44.7	60.9	47.8	56.2
Hossian & Little et al.[202]	66.1	59.3	72.6	57.3	46.6	62.4	46.6	58.3
Yang et al.[13]	69.2	57.4	65.4	58.4	60.1	43.6	47.7	58.6
Ours	59.1	46.2	53.9	46.1	41.2	48.1	36.4	45.8

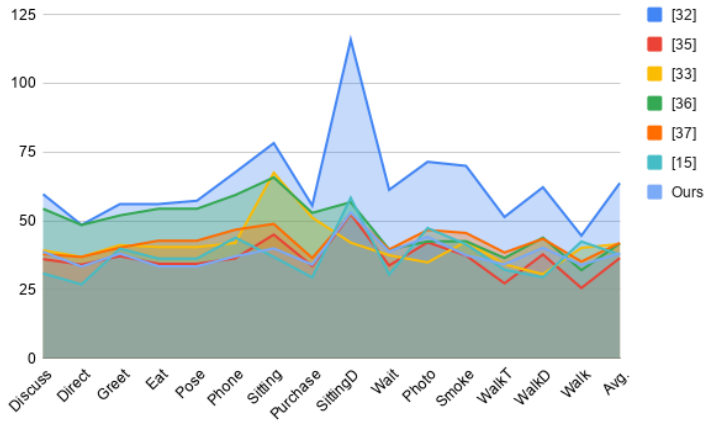

FIGURE 5.10: Early and Late fusion network over Heatmap and Frame Flow

we predict 2D joint heatmaps using the proposed ESHA, which is based on the stack-hourglass method. The resulted heatmaps have been fused with the original image in

TABLE 5.3: The Evaluation Criteria 2 over Human3.6M public data set has been employed to calculate the 3D HPR errors (in the form of millimeters).

Human 3.6M								
Algorithm	**Greet**	**Direct**	**Discuss**	**Phone**	**Pose**	**Eat**	**SittingD**	**Purchase**
Criteria 2								
Wang et al. [207]	56.12	48.54	59.71	67.68	57.31	56.12	115.85	55.57
Pavllo et al.[142]	37.2	34.1	36.1	36.4	34.4	34.4	52.5	33.6
Chen et al. [196]	41.2	36.9	39.3	42.0	38.0	40.5	42.1	51.2
Pavlakos et al.[200]	52.0	48.5	54.4	59.4	49.9	54.4	56.8	52.9
Hossian & Little et al. [202]	40.3	36.9	37.9	46.8	37.7	42.8	52.6	36.5
Yang et al.[13]	39.9	26.9	30.9	43.9	28.8	36.3	58.4	29.4
Ours	38.8	33.4	38.1	33.5	36.9	37.1	40.0	34.1

	Sitting	**Smoke**	**Photo**	**Wait**	**Walk**	**WalkDog**	**WalkT**	**Average**
Wang et al. [207]	78.26	69.99	71.47	61.29	44.63	62.22	51.42	63.74
Pavllo et al.[142]	45.0	37.4	42.2	33.8	25.6	37.8	27.3	36.5
Chen et al.[196]	67.5	42.5	34.9	37.5	40.2	30.6	34.2	41.6
Pavlakos et al.[200]	65.8	42.6	42.5	39.6	32.1	43.9	36.5	41.8
Hossian & Little et al. [202]	48.9	45.6	46.7	39.6	35.2	43.5	38.5	42.0
Yang et al.[13]	36.9	41.5	47.4	30.5	42.5	29.5	32.2	37.7
Ours	53.0	38.9	44.2	37.5	34.2	40.2	34.0	38.26

TABLE 5.4: The Evaluation Criteria 3 over Human3.6M public data set has been employed to calculate the 3D HPR errors (in the form of millimeters).

Human 3.6M								
Algorithm	**Greet**	**Direct**	**Discuss**	**Phone**	**Pose**	**Eat**	**SittingD**	**Purchase**
Criteria 3								
Wang et al. [207]	48.90	38.69	45.62	77.3	47.49	54.77	94.30	47.17
Chen et al.[196]	50.8	45.9	48.0	48.9	46.1	48.6	49.4	64.73
Pavlakos et al.[200]	89.9	79.2	85.2	86.3	75.8	78.3	137.6	81.8
Ours	46.7	33.2	42.8	40.9	44.0	48.3	57.5	43.7

	Sitting	**Smoke**	**Photo**	**Wait**	**Walk**	**WalkDog**	**WalkT**	**Average**
Wang et al. [207]	54.65	56.84	78.85	49.29	33.07	58.71	38.96	54.14
Chen et al.[196]	57.4	54.2	45.1	47.2	49.9	39.9	42.9	50.3
Pavlakos et al.[200]	106.4	86.2	87.9	92.3	82.3	72.9	77.5	88.6
Ours	48.9	57.4	44.0	55.0	43.9	46.1	38.5	45.9

multi-view senario for 3D reconstruction. The novelty of the method lies in the introduction of ESHA method for 2D pose and utilization of both the data information in multi-view setup. Experimental outcomes prove that the proposed design choice of the deep network can significantly enhance the performance on the Human3.6M dataset in terms of MPJPE and PA-MPJPE error metric. Through these experimental outcomes,

the authors also proved that utilizing a fusion of the 2D joint heatmap and image information in the multi-view scenario enhances the 3D HPR functioning in the form of the reconstruction error. In the future job, authors intend to examine this type of fusion to reconstruct 3D HPR in-the-wild videos.