Multi-Person 3D Human Pose Estimation from Monocular Images

Rishabh Dabral
IIT Bombay
rdabral@cse.iitb.ac.in

Nitesh B Gundavarapu
IIT Bombay
ntesh93@gmail.com

Rahul Mitra
IIT Bombay
rmitter@cse.iitb.ac.in

Abhishek Sharma
Axogyan AI
abhisharayiya@gmail.com

Ganesh Ramakrishnan
IIT Bombay
ganesh@cse.iitb.ac.in

Arjun Jain
Axogyan AI
arjunjain@gmail.com

Abstract

Multi-person 3D human pose estimation from a single image is a challenging problem, especially for in-the-wild settings due to the lack of 3D annotated data. We propose HG-RCNN, a Mask-RCNN based network that also leverages the benefits of the Hourglass architecture for multi-person 3D Human Pose Estimation. A two-staged approach is presented that first estimates the 2D keypoints in every Region of Interest (RoI) and then lifts the estimated keypoints to 3D. Finally, the estimated 3D poses are placed in camera-coordinates using weak-perspective projection assumption and joint optimization of focal length and root translations. The result is a simple and modular network for multi-person 3D human pose estimation that does not require any multi-person 3D pose dataset. Despite its simple formulation, HG-RCNN achieves the state-of-the-art results on MuPoTS-3D while also approximating the 3D pose in the camera-coordinate system.

1. Introduction

3D human pose-estimation consists of inferring the 3D joint-locations from an image or a sequence of images. It is the key to unlocking a large number of applications in AR/VR, Human-Computer-Interaction (HCI), Gaming, Activity Recognition, Surveillance, *etc.*. Although, there is a vast literature on single-person 3D pose estimation [31, 13, 40, 41, 24, 21, 5, 36, 3, 16, 1, 39, 6, 12], the space of multi-person 3D pose estimation is mostly unexplored with only a handful of prior work [27, 20, 37, 28, 38]. Ironically, real-life human pose-estimation applications, most often, require multi-person pose estimation. For example, surveillance systems require real-time capturing of the poses for every person in the scene. Similarly, sports-analytics demands that all the players are simultaneously analyzed to capture inter-player interactions. Consequently, there ex-

ists a gap between existing research and real-world requirements.

A simple extension of the single-person pose estimation systems to the multi-person setting involves separate detection of every person followed by single-person pose estimation on person crop.

Unfortunately, the run-time of this approach is likely to increase linearly with the number of people in the scene, making it inefficient for analysis in crowded scenes. Additionally, most existing multi-person pose estimation methods [27, 20, 28], with the exception of [37] estimate 3D pose configuration only relative to the root joint. However, relative spatial ordering of different people in the scene is also needed to facilitate reasoning about human interactions and provide a better understanding of the scene. Relative spatial estimation has the potential to unlock accurate tracking of multiple persons in a scene video.

Moreover, most prior work on multi-person pose estimation [27, 20, 38] relies on creating or simulating a multi-person 3D human pose dataset as a necessity for training. The pre-requisite is due to the end-to-end integrated person detection and pose estimation pipeline. This limits the variability presented to the system while training because obtaining real-world in-the-wild 3D annotations in multi-person setting is challenging, expensive and a research problem in itself.

In light of the aforementioned discussion of multi-person 3D pose estimation, we propose a quasi top-down architecture that decouples the 2D key-point detection and 2D-to-3D lifting tasks. The proposed architecture, HG-RCNN, brings together the goodness of Mask-RCNN [9] and the Hourglass [23] network for heatmap regression. The regressed heatmaps are then fed to an independently trained lifting module to regress the root-relative 3D poses. Consequently, we completely avoid using any multi-person 3D pose dataset in the pipeline since it leverages the existing multi-person 2D pose datasets and single-person 3D

Figure 1. Some results of our proposed 3D pose estimation pipeline on some challenging samples from MS COCO. Our approach is resilient against occlusions and clutter. We also approximate the spatial ordering of people in the scene with respect to the camera. Further in-the-wild results and a 3D rendered view of the above images can be found in the supplementary material.

pose datasets. Owing to its modular architecture, the first step of obtaining 2D poses can be trained with publicly available large-scale in-the-wild multi-person datasets, such as COCO [15], LIP [7] and MPII 2D dataset [2]. This allows HG-RCNN to cope with challenging variations in view-point, lighting, apparel, occlusion and extreme poses without the need of costly 3D annotations in-the-wild setting. The keypoint heatmaps from the HG-RCNN are passed through a *soft-argmax* module and fed to a 2D-3D lifting module. Finally, our pipeline approximates pose-configurations in camera coordinates without the need of costly geometric optimization. The resulting system outperforms all previous approaches on the challenging MuPoTS-3D [20] test-set that contains a majority of in-the-wild test scenarios. The method generalizes well to in-the-wild images, even without exploiting any structural priors, while running at 12-15fps on images of size 400×600 on a single Nvidia 1080Ti graphics card.

In summary, we contribute a state-of-the-art model for performing in-the-wild multi-person 3D pose estimation. The model can be trained without using any multi-person 3D dataset and the system also estimates the relative ordering of the persons in the 3D space.

2. Related Works

Human Pose Estimation has been a widely studied problem. Here, we describe prior art relevant to this work from three broad viewpoints: (a) 2D Pose estimation, (b) Single-person 3D Pose estimation and (c) Multi-Person 3D Pose estimation. A detailed survey of the area can be found in [29].

2D Human Pose Estimation: Most 2D human pose estimation methods represent their joint outputs as heatmaps, wherein a heatmap's value at a point represents the possibility of the corresponding joint's existence in that position. [34] proposed Convolutional Pose Machines that iteratively refined the heatmap predictions at every stage. The Stacked Hourglass network [23] was an encoder-decoder architecture with skip connections to facilitate joint reasoning of high level structural and low level textural features of human pose. Mask-RCNN [9] proposed an extension of Faster-RCNN [26] for simultaneously predicting the pose and 2D keypoints and/or instance segmentation masks. In a similar line of work, [8] predicted the *u-v* maps of the persons which can then be used for dense reconstruction. [30] proposed a variant to Mask-RCNN by defining joints as regions instead of persons. In similar spirits, our proposed pipeline attempts to synergise Mask-RCNN and Hourglass networks for multi-person 3D pose estimation task.

Single-person 3D Pose Estimation: Single person 3D pose estimation works can be broadly divided based on whether they directly regress 3D joints [31, 13, 40, 41] or use a pipelined approach of inferring 3D pose from 2D pose [33, 39, 21, 22, 14]. VNect [21] proposed the first real-time approach and parameterized a 3D joint by a heatmap and 3 location maps. Using a 2D-to-3D pipeline enables the use of rich 2D pose datasets which, in turn, improves in-the-wild generalizability. Many approaches perform a direct 2D-to-3D lifting of poses [41, 17, 22, 5, 36] by either learning the transformation or by a nearest-neighbour lookup in a pose library. Furthermore, many pipelined approaches [21, 27, 39, 31, 41, 24] have reported significant improvements in in-the-wild performances by using the more diverse 2D pose datasets to pre-train or jointly train their 2D prediction modules.

Several methods in the past have also reported significant improvements by using temporal cues [25, 21, 41, 6, 35, 37] by either learning a motion/refinement model or by using temporal constraints in a constrained optimization framework.

Multi-Person 3D Pose Estimation: Broadly, multi-person pose estimation approaches, 2D and 3D alike, can be classified into top-down and bottom-up approaches. Bottom-up approaches simultaneously predict all the key-points followed by assembling them into full poses for all persons. On the other hand, top-down approaches first detect the human candidates and subsequently perform pose estimation for each of them. While bottom-up methods are lucrative in terms of efficiency, they tend to be less accurate. For example, the top 5 entries in MS COCO key-points challenge employ top-down approaches [15]. Intuitively, it makes sense to solve for pose estimation on a person's crop, instead of solving a much more challenging problem of grouping detected key-points into a full person. In recent years, however, a middle ground has been found in the form of quasi top-down architectures based on Mask-RCNN [9, 8, 10, 30] that have been successful in simultaneously detecting the object RoIs and performing downstream tasks on the corresponding RoI feature-maps, without having to crop the image back.

LCRNet [27] was the first method to perform Multi-Person 3D Pose Estimation. They propose an integrated network based on Faster-RCNN [26] which first proposes Regions of Interest (RoIs) that are fed to a classifier and a regressor. The classifier estimates the most probable anchor pose out of the K pre-defined anchor poses obtained from a MoCap dataset. The regressor then refines the anchor poses towards an accurate pose prediction. Alternately, [20] propose a bottom-up approach wherein they regress the heatmaps along with X, Y, and Z location maps for every image. The location maps provide the corresponding 3D positions of joints in metric space. The estimated 3D joints

are then associated using Part Affinity fields [4] based on the heatmaps. Both the approaches depend on the explicit creation or simulation of multi-person 3D pose datasets for training. Our method, on the other hand, avoids the use of such datasets and relies on 3D data only for the single person case. Further, Zanfir *et al*. [37] proposed a large-scale human sensing system for multiple people that estimates pose and shape using the top-down approach of person detection followed by pose estimation for each person. Recently, Zanfir *et al*. [38] proposed MubyNet, a bottom-up approach that performs joint association by formulating it as a binary integer programming problem. In contrast, Mask-RCNN [9] based quasi top-down methods [8, 10, 27] have proven to be effective for simultaneously locating objects at a coarse level and detecting finer spatial layouts like segmentation masks, key-point heatmaps, u-v maps, etc.. Our proposed HG-RCNN exploits this setting and also regresses for 3D key-points. However, unlike LCRNet [27] and LCR-Net++ [28], our method does not require anchor-poses and is relatively simpler.

3. Problem Formulation

Given an image I containing N people, we estimate the poses $P = (P_1, P_2, ..., P_N)$, wherein $P_i \in \mathcal{R}^{n \times 3}$ and n is the number of joints. Every pose P_i is a set of n joints in 3D Euclidean space with the origin set to a root joint, pelvis in this case. As an intermediate step, our method first estimates the 2D key-points $K = (K_1, K_2, ..., K_N)$ with $K_i \in \mathcal{R}^{n \times 2}$ in the image coordinate space. Finally, we approximate the global poses $P^G = (P_1^G, P_2^G, ..., P_N^G)$ in camera coordinate space.

3.1. Multi-Person 3D Pose Estimation

We follow a generic, two-step pipeline for root-relative 3D pose estimation. First, we estimate per-frame 2D key-points of all the people in an image and lift them to 3D pose using a simple residual network. We use a Mask-RCNN based architecture to estimate 2D key-points. However, vanilla keypoint head of Mask-RCNN is not the most conducive architecture for reasoning with structured/articulated objects like human pose. Fortunately, the Hourglass [23] family of networks have been found to be extremely effective in reasoning about a human pose in a structure-aware way. Therefore, we propose to employ a tiny Hourglass head as a surrogate to the key-point head. This simple patch alone leads to noticeable improvements in the results and will be discussed further in Section 5.2.

In the second step, the obtained keypoint heatmaps are lifted to 3D joints using a network with two residual modules of size 2048. When deployed in wild settings, it is trained with the heatmaps regressed on the MPI-INF-3DHP training dataset [18] which provides a wide variety of viewpoints and poses activities, thereby adding to the general-
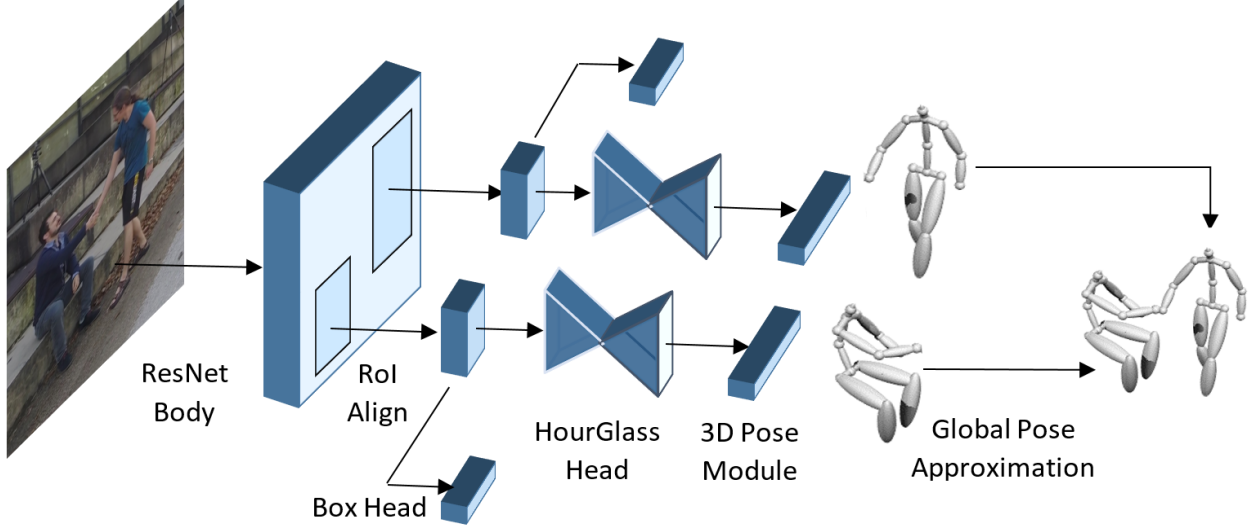
Figure 2. Schematic of our Multi-Person 3D Pose Estimation approach. We augment the Faster-RCNN [26] architecture with a shallow HourGlass Network [23]. The heatmaps generated by the hourglass are then input to a 3D Pose Module which regresses the root-relative 3D joint coordinates. The estimated 3D poses of all the Regions of Interests (RoI's) are then collected and their global root positions are approximated to ensure that relative spatial ordering is preserved.

ization capability of the network. It is worth noting that it is this modular structure of the pipeline that allows us to train the network without any multi-person 3D dataset. The in-the-wild performance is guaranteed by two aspects: a) The heatmaps are learnt on completely wild multi-person 2D keypoint datasets, and b) the lifting module is agnostic to the image features and trained on a dataset consisting of a wide variety of 2D-3D paired annotations.

Further details on the architecture are discussed in Section 4. At this stage, all the outputs (3D keypoints) are in their individual root relative space. For placing the detected poses in camera-relative space, we estimate the common focal length of the camera and the translation vectors from the individuals' roots to the camera center.

3.2. Global Pose Approximation

Our approach for camera-relative pose approximation is based on jointly optimizing the root joints' global positions and the camera's focal length for the projection error. We initialize the root joint positions using a *weak-perspective* projection assumption, thus, requiring us to estimate the shrinking parameter α_i for every pose Pi in the scene. To this end, we compute the sum of bone lengths of the 2D keypoints, S_{2D}, followed by computing the sum of bone lengths, S_{3D}, of the 3D pose's orthographic projection.

The ratio S_{2D}/S_{3D} acts as a surrogate to the shrinking factor α_i. This finally leads to the following formulation for estimating the global X (horizontal) and Z (depth) coordinates of a joint:

$$Z = f * \frac{S_{3D}}{S_{2D}}, \qquad (1)$$

$$X = (x - o_x) * \frac{S_{3D}}{S_{2D}} \qquad (2)$$

where, x corresponds to the 2D keypoint and o_x is the x co-ordinate of the image center. The focal length, f, is initialized by assuming a field-of-view of $60°$. The same formulation holds for the Y (vertical) coordinate as well.

Once the root translations are initialized and the full 3D poses are placed in the respective root positions, we iteratively optimize the translation and focal length. The global rotations are assumed to be identity. Thus, the objective function can be written as:

$$f^*, t^* = \arg \min_{f,t} \sum_{i=1}^{N} ||K_i - \Pi_{f,t_i} P_i||_2 \qquad (3)$$

where $t = \{t_1, t_2, \ldots t_N\}$ with t_i being the translation vector of ith subject's root joint and Π being the projection operator. This, finally, leads to the global pose, $P_i^G = P_i + t_i^*$.

It is worth noting that the proposed global pose approximation method is just an approximation that can be quickly implemented and run in real-time. The approximation is not expected to work when the person is aligned with the optical axis. We discuss further limitations in section 6. It is not intended to be highly accurate, but only expected to make spatial ordering apparent to systems that need it, eg. action recognition.

4. Network and Training Details

HG-RCNN: The HG-RCNN is constructed by appending an hourglass on the keypoint head of Mask RCNN as shown in Figure (2). Instead of upsampling once while deconvolving and once at the final layer, we upsample (with

Table 1. Comparison of our method with prior work on MuPoTS-3D on *Setting 1*. The **top half** shows results on *all annotated poses* in the test set. The **bottom half** shows results when only the detected poses are considered. The evaluation metric is 3D PCK and higher is better. *Note, that the average PCK provided in LCRNet++ [28] is not weighed by the number of persons in each test sequence unlike [27, 20] and ours.

Method	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20	Avg
[27]	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
[20]	81.0	59.9	64.4	62.8	68.0	30.3	65.0	59.2	64.1	83.9	67.2	68.3	60.6	56.5	69.9	79.4	79.6	66.1	66.3	63.5	65.0
[28]*	87.3	61.9	67.9	74.6	**78.8**	48.9	58.3	59.7	**78.1**	**89.5**	69.2	**73.8**	**66.2**	56.0	**74.1**	82.1	78.1	72.6	73.1	61.0	70.6
[19]	**88.4**	65.1	68.2	72.5	76.2	46.2	**65.8**	**64.1**	75.1	82.4	74.1	72.4	64.4	**58.8**	73.7	80.4	**84.3**	67.2	74.3	67.8	70.4
Ours	85.1	**67.9**	**73.5**	**76.2**	74.9	**52.5**	65.7	63.6	56.3	77.8	**76.4**	70.1	65.3	51.7	69.5	**87.0**	82.1	**80.3**	**78.5**	**70.7**	**71.3**
[27]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
[20]	81.0	64.3	64.6	63.7	73.8	30.3	65.1	60.7	64.1	83.9	71.5	69.6	69.0	69.6	71.1	82.9	79.6	72.2	76.2	85.9	69.8
[28]*	88.0	73.3	67.9	**74.6**	81.8	50.1	60.6	60.8	**78.2**	**89.5**	70.8	**74.4**	**72.8**	64.5	**74.2**	84.9	85.2	78.4	75.8	74.4	74.0
[19]	**88.4**	70.4	**68.3**	73.6	**82.4**	46.4	66.1	**83.4**	75.1	82.4	76.5	73.0	72.4	**73.8**	74.0	83.6	84.3	73.9	**85.7**	90.6	**75.8**
Ours	85.8	**73.6**	61.1	55.7	77.9	**53.3**	**75.1**	65.5	54.2	81.3	**82.2**	71.0	70.1	67.7	69.9	**90.5**	**85.7**	**86.3**	85.0	**91.4**	74.2

Table 2. Performance of our method on MuPoTS using the *Setting 2*. The **top half** shows results on all annotated poses in the test set. The **bottom half** shows results when only the detected poses are considered. 'all' corresponds to evaluation on all eligible persons and 'occ' corresponds to the results on occluded persons. The evaluation metric is 3D PCK and higher is better. Notice that compared to 1, the improvement is mostly observed on sequences with significant occlusion, eg. TS18 and TS19.

Method	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20	Avg
Ours (occ)	82.0	61.1	62.3	70.2	62.7	53.3	67.8	63.1	59.8	39.1	73.1	69.3	67.4	33.7	59.0	79.1	79.0	82.5	79.7	36.2	64.0
Ours (all)	85.2	69.2	73.1	75.6	77.7	52.1	65.1	66.0	57.7	77.6	76.6	69.0	71.6	53.4	70.3	86.0	84.3	84.7	83.7	72.8	72.6
Ours (occ)	82.0	61.1	62.3	71.0	62.7	53.8	69.0	63.2	59.8	39.1	78.0	69.8	67.4	47.4	59.0	79.2	79.5	82.5	79.7	76.6	67.1
Ours (all)	85.2	69.2	73.1	76.1	77.7	52.7	65.9	66.0	57.7	77.6	80.5	69.1	71.6	60.6	70.3	87.1	85.1	84.7	83.7	92.1	74.3

$4\times$) the feature maps all at once before passing the feature values on to the hourglass. The number of feature-maps is brought down from 512 to 128 using a 1×1 convolution layer. The original hourglass is modified to have three nested residuals (instead of 4) and has a feature-map of size 7×7 at the bottle-neck layer. The hourglass output is then fed to a final classification layer which predicts the heatmaps for every joint.

We train the network described above with the Cross-Entropy Loss. While finetuning, we train on top 500 RoIs and use a batch size of 16. The network is trained with a base learning rate of 0.02 on a single Nvidia P6000 Quadro graphics card.

3D Pose Module: Our 2D-to-3D pose module converts the heatmap activations to 3D pose using a residual architecture and is in line with the 2D-3D lifting pipelines proposed in [17, 32, 22]. We input the 2D poses in heatmap space after passing the heatmaps through a *softargmax* layer. This has two benefits: a) it makes learning possible from images of any given size and scale, and b) it facilitates end-to-end training of the network architecture. The network is trained using RMSProp optimizer and a learning rate of $2.5 \exp -4$ which is reduced by 10 times after 40 epochs.

While testing on MuPoTS (multi-person) dataset, we use the 3D pose module trained only on MPI-INF-3DHP dataset because both the training and the test sets had the same motion capture system. Human3.6 was captured by a different mocap system which leads to the same joint name pointing to different physical locations on the body.

5. Experiments

This section describes our experiments on MuPoTS-3D [20], MS COCO [15] and Human3.6M [11] datasets.

5.1. Evaluation Datasets

MuPoTS-3D Test Set: Multi-Person Test Set 3D [20] is a recently released *multi-person* 3D human pose test dataset. It consists of 20 test sequences shot with a marker-less mocap system - 5 indoor and 15 outdoor. Every sequence contains 2-3 persons in a variety of activities. The evaluation metric used is 3D PCK - percentage of correct keypoints within a radius of 15cm - on all the annotated persons. In case of a missed detection, all the joints of the missed person are considered erroneous. An alternative evaluation mode is the one in which the evaluations are performed only on the detected joints.

The official evaluation code performs a greedy matching of detections and ground truth based on the number of 2D keypoints within a proximity of $40px$. We call this method *Setting 1* for MuPoTS.

We also evaluate our model in the setting wherein the greedy matching is done based on 3D distances instead of 2D distances. We call this *Setting 2*. This joint matching strategy is, arguably, less sensitive to cases of heavy oc-

Figure 3. Visualization of our results on MuPoTS-3D Test Set from different viewpoints. Notice that the model is fairly robust to occlusions. The spatial alignment is not derived from ground truth.

clusion which would, otherwise, confuse a keypoint based matching detector. This, as discussed in Section 5.2, leads to missed detections even when the model actually detects the appropriate person. Note, that the two settings differ only in the way the predicted poses are matched with the ground truth poses. All the other details of evaluation, like 3D PCK threshold, joints used for matching, etc remains the same.

Human3.6: Human 3.6M [11] is a single-person 3D human pose dataset captured with marker-based motion capture system. It consists of 11 subjects performing 15 actions. We evaluate our model on the commonly followed protocol [21, 31, 39, 27, 18, 6, 22] that uses subjects $1, 5, 6, 7$ and 8 for training, The evaluations are done on subjects 9 and 11. All the videos are downsampled from $50 fps$ to $10 fps$. The evaluation metric used is Mean Per Joint Position Error (MPJPE) which is calculated after aligning only the roots of the predicted and ground truth 3D poses.

MSCOCO Keypoints: MSCOCO Keypoints is a large scale dataset for 2D multi-person keypoint detection task with roughly 110k training images. It also provides the person bounding boxes and segmentation masks. The 2D keypoint detection task is evaluated on the commonly used Average Precision (AP) metric at different threshold levels. Similarly, the quality of bounding box detections are evaluated using AP.

5.2. Quantitative Evaluation

We now discuss the numerical results achieved on the datasets mentioned above.

MuPoTS-3D Test Set: Table 1 compares the performance of our simple yet effective method with the existing multi-person 3D pose results. On *Setting 1*, we improve the state-of-the art significantly with a 3DPCK of 71.25% as against 65% in [20] and 53.8% in [27]. For LCRNet [27], the reported results are evaluated by [20]. We report an improved performance on several test sequences. We also significantly improve the performance of occluded joints (61% vs 48.7%) as well as the non-occluded joints (75.6% vs 70%) when compared with [20]. Our method also performs significantly well when only detected persons are compared. In this setting, we observe 75% 3DPCK while the state-of-the-art being 69.8%. We also compare our performance with the recently released XNect [19] and demonstrate competitive results on all annotated poses (71.3% vs 70.4%) as well as the detected poses (74.2% vs 75.8%).

We also evaluate our method on the proposed *Setting 2*. We observed an improved 3DPCK of 72.6% when compared with *Setting 1*. This improvement is facilitated by a simple tweak in the greedy matching algorithm of ground-truth and predicted persons. On deeper inspection, we see sharp improvements in sequences with heavy occlusions, like TS18 and TS19. Further, the overall improvement is significant when comparing the performance of occluded joints (64% vs 61% of [20]). This observation can be attributed to the fact that matching predictions with ground-

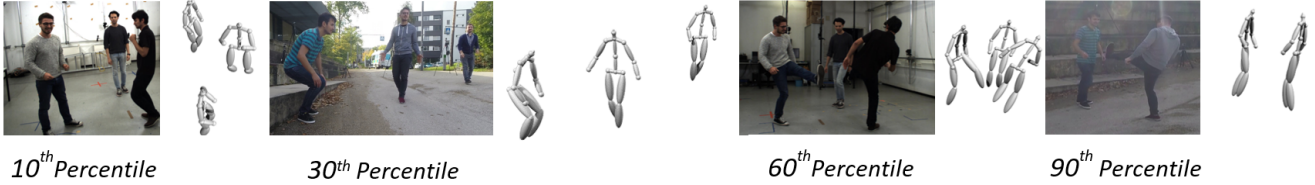10th Percentile 30th Percentile 60th Percentile 90th Percentile

Figure 4. MPJPE based Percentile Analysis on MuPoTS-3D test set. The lower percentile is better. An important inference from the analysis is that the method is sensitive to lighting and low contrast setting.

Table 3. Comparative evaluation of our model on Human 3.6 using Absolute MPJPE. The evaluations were performed on subjects 9 and 11. The papers above the horizontal line are single-person pose estimation papers and the ones below the line are multi-person pose estimation papers.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Martinez [17]	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0
Zhou [39]	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2
Sun [31]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7
Dabral [6]	44.8	50.4	44.7	49.0	52.9	43.5	45.5	63.1
Hossain [25]	44.2	46.7	52.3	49.3	59.9	47.5	46.2	59.9
Sun [32]	47.5	47.7	49.5	50.2	51.4	43.8	46.4	58.9
Rogez [27]	76.2	80.2	75.8	83.3	92.2	79.0	71.7	105.9
Mehta [20]	58.2	67.3	61.2	65.7	75.8	62.2	64.6	82.0
Rogez [28]	50.9	55.9	63.3	56.0	65.1	52.1	51.9	81.1
Ours (Baseline)	60.2	64.5	66.2	70.1	75.6	65.4	69.4	83.7
Ours (Fine-Tuned)	52.6	61.0	58.8	61.0	69.5	58.8	57.2	76.0
Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
Martinez [17]	94.6	62.3	78.4	59.1	65.1	49.5	52.4	62.9
Zhou [39]	111.6	64.1	65.5	66.0	51.4	63.2	55.3	64.9
Sun [31]	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Dabral [6]	87.3	51.7	61.4	48.5	37.6	52.2	41.9	52.1
Hossain [25]	65.6	55.8	59.4	50.4	52.3	43.5	45.1	51.9
Sun [32]	65.7	49.4	55.8	47.8	38.9	49.0	43.8	49.6
Rogez [27]	127.1	88.0	105.7	83.7	64.9	86.6	84.0	87.7
Mehta [20]	93.0	68.8	84.5	65.1	57.6	72.0	63.6	69.9
Rogez [28]	91.7	64.7	70.7	54.6	44.7	61.1	53.7	**61.2**
Ours (Baseline)	105.7	70.2	89.6	69.1	61.7	80.6	66.9	73.0
Ours (Fine-Tuned)	93.6	63.1	79.3	63.9	51.5	71.4	53.5	65.2

truths based on 2D keypoints leads to matching errors and missed detections when two or more persons occlude each other. Indeed, we observe that the algorithm's detection percentage rose from 93% to 96%, thus improving the overall 3DPCK. Interestingly, we observe that TS10 suffers under this protocol because all the three subjects bear similar poses for many frames. Thus, we believe the two settings are complimentary. Another evaluation metric used in MuPots Test Set is the Area Under Curve (AUC) of PCK values. We report an AUC of 35.5 which is better than 30.1 reported by [20] and 27.6 in [21] using groud truth detections. Our detection rate is 93.5% which is comparable to the 93% detection rate of [20] under *Setting 1*.

The above mentioned results reveal a significant increment in the state-of-the-art. It is worth noting that all the results are comparable to performance of single-person pose estimation methods.

Table 4. Performance comparison of various training/testing settings on Human3.6M Protocol 1. The first column indicates the data used as the 2D input to the 3D pose module while training. The second column, likewise, indicates which datasets were used for training the HG-RCNN based 2D input.

2D-3D Training	HG-RCNN Training	MPJPE
H36M GT	MS-COCO	135.5
H36M GT + noise	MS-COCO	119.7
H36M pred	MS-COCO	73.0
H36M pred	MS-COCO + H36M	65.2
MPI-INF GT	MS-COCO	118.3
MPI-INF GT + noise	MS-COCO	118.16

Human 3.6M: The results on Human 3.6M dataset are detailed in Table 3. We achieve an MPJPE of $65.2mm$ after fine-tuning HG-RCNN on Human3.6M and $74.3mm$ without fine-tuning. It may be noted that Zanfir et al. [38] report their results on the official Human3.6 test set and achieve

$60mm$ MPJPE. Since the test circumstances are different, the comparison may not be fair. The combined results on MuPoTS-3D and Human3.6M also corroborate the claims in [12] that a good performance in Human3.6M does not necessarily indicate better generalization in wild settings. We also evaluate our method under various test-train settings in Table. 4 and observe that MPI-INF-3DHP [18] offers a wider range of poses to train from, thus leading to better results with ground-truth detections.

Table 5. Comparison of HG-RCNN and Mask-RCNN based models on MuPoTS 3D. The evaluation metric is 3DPCK.

	Mask-RCNN	HG-RCNN
all annotated joints	70.1	72.4
all occluded joints	61.0	64.1

Mask-RCNN vs. HG-RCNN: Table 6 details the performance of HG-RCNN on MSCOCO Keypoints dataset. Our results are comparable to Mask-RCNN's reported results. We observe a slightly reduced mAP which can be attributed to the fact that Hourglass architecture is better suited for cases when the larger structure is to be considered. MS-COCO keypoints validation dataset contains multiple cases of isolated/truncated body parts. While evaluating on MuPoTS-3D, we observe improved 3DPCK using HG-RCNN on all annotated (70.1% vs 72.4%) and all occluded (61% vs 64%) joints alike. We also achieve comparable results on the person bounding box detections over Mask-RCNN as shown in Table 7.

Table 6. HG-RCNN results on MS-COCO 2017 *val-set* for keypoints using a ResNeXt-101 backbone.

	AP	AP50	AP75	AP_M	AP_L
HG-RCNN	0.6348	0.8620	0.6905	0.5840	0.7204

Table 7. Results on MS-COCO 2017 *val-set* for person boxes.

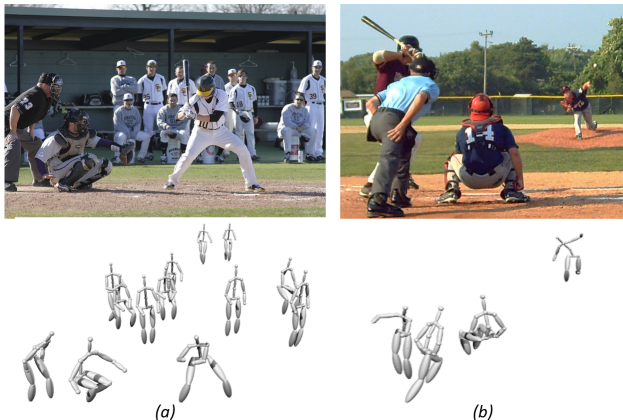|	AP	AP50	AP75	AP_S	AP_M	AP_L
HG-RCNN	0.5536	0.8381	0.6076	0.3743	0.6320	0.7235

Figure 5. Two images summing up the sources of failure in our approach.

6. Limitations

While our method attempts to account for structural information during inter-personal occlusions, we believe it can be explicitly taken care of with better structural constraints and bounding box consistencies.

Sources of Error: Figure 5 shows interesting examples of failure cases and exposes three sources of error in our pipeline. The first source is poor 2D keypoint estimation, which is apparent in the occluding persons of Figure 5 (b). The second source of error is an unseen activity/pose which leads to erroneous prediction. This can be seen in squatting players of both the figures, wherein the data-induced model bias leads to incorrectly predicting a person sitting on a chair instead.

Finally, our camera-coordinate 3D pose prediction is sensitive to 2D keypoint detections and can wrongly reason about the person depth. This effect is observable in Figure 5(a) in which the sitting people have been pushed back, in addition to the two outliers standing behind the player. It may also be noted that this approximation also assumes the individuals to be of roughly the same size. We observe incorrect relative positioning when the height difference is high. Finally, while we compute the sums of bone lengths only on the torso joints to avoid the adverse effects of foreshortening, the effects can not be completely alleviated.

7. Conclusion

This paper presents a simple extension of Faster-RCNN framework to yield a near-real-time multi-person 3D human pose estimation network HG-RCNN that can be trained without a multi-person 3D pose dataset. Our proposed framework is extremely simple to implement and outperforms previous state-of-the-art results by convincing margins. We also show that we can approximate the spatial layout of the scene. These claims are substantiated both quantitatively through experimental evaluation as well as through qualitative assessments on COCO and MuPoTS-3D datasets. The paper also proposes an improvement to the greedy-matching strategy for multi-person 3D pose estimation evaluation and show results on it. In the future, we plan to deploy this pipeline to a broader human-parsing pipeline while also seeking real-life applications such as activity detection and construct a better scene understanding system related to humans.

Acknowledgement

References

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, 2015. 1

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2

[3] F. Bogo, 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. In *ECCV*, 2016. 1

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[5] C.-H. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *CVPR*, 2017. 1, 3

[6] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain. Learning 3d human pose from structure and motion. In *ECCV*, 2018. 1, 3, 6, 7

[7] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 2

[8] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 3

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 2, 3

[10] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick. Learning to Segment Every Thing. In *CVPR*, 2018. 3

[11] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2014. 5, 6

[12] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 1, 8

[13] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 1, 3

[14] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng. Recurrent 3d pose sequence machines. In *CVPR*, 2017. 3

[15] T. Lin, 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. *arXiv preprint arXiv:1405.0312*, 2014. 2, 3, 5

[16] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 1

[17] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 3, 5, 7

[18] D. Mehta, 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. In *3DV*, 2017. 3, 6, 8

[19] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. *arXiv preprint arXiv:1907.00837*, 2019. 5, 6

[20] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 1, 2, 3, 5, 6, 7

[21] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In *ACM ToG*, 2017. 1, 3, 6, 7

[22] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017. 3, 5, 6

[23] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 2, 3, 4

[24] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017. 1, 3

[25] M. Rayat Imtiaz Hossain and J. J. Little. Exploiting temporal information for 3d human pose estimation. In *ECCV*, 2018. 3, 7

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 3, 4

[27] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net: Localization-classification- regression for human pose. In *CVPR*, 2017. 1, 3, 5, 6, 7

[28] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *TPAMI*, January 2019. 1, 3, 5, 7

[29] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 2016. 2

[30] T. Sekii. Pose proposal networks. In *ECCV*, 2018. 2, 3

[31] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, 2017. 1, 3, 6, 7

[32] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *ECCV*, 2018. 5, 7

[33] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *CVPR*, 2017. 3

[34] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *CVPR*, 2016. 2

[35] B. Xiaohan Nie, P. Wei, and S.-C. Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *ICCV*, Oct 2017. 3

[36] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A dual-source approach for 3d pose estimation from a single image. In *CVPR*, 2016. 1, 3

[37] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *CVPR*, 2018. 1, 3

[38] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, and C. Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *NeurIPS*. 2018. 1, 3, 7

[39] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *ICCV*, 2017. 1, 3, 6, 7

[40] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCV Workshops*, 2016. 1, 3

[41] X. Zhou, M. Zhu, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, 2016. 1, 3