Single-shot 3D multi-person pose estimation in complex images

Abdallah Benzine, Bertrand Luvison, Quoc-Cuong Pham, Catherine Achard

Single-shot 3D multi-person pose estimation in complex images

Abdallah Benzine, Bertrand Luvison, Quoc-Cuong Pham, Catherine Achard

Single shot 3D multi-person pose estimation in complex images

Abdallah Benzine[⋆,†], Bertrand Luvison[⋆], Quoc Cuong Pham[⋆], Catherine Achard [†]

[⋆] *CEA, LIST, Vision and Learning Lab for Scene Analysis, PC 184, F-91191 Gif-sur-Yvette, France*

[†] *Sorbonne University, CNRS, Institute for Intelligent Systems and Robotics, ISIR, F-75005 Paris, France*

Abstract

In this paper, we propose a new single shot method for multi-person 3D human pose estimation in complex images. The model jointly learns to locate the human joints in the image, to estimate their 3D coordinates and to group these predictions into full human skeletons. The proposed method deals with a variable number of people and does not need bounding boxes to estimate the 3D poses. It leverages and extends the Stacked Hourglass Network and its multi-scale feature learning to manage multi-person situations. Thus, we exploit a robust 3D human pose formulation to fully describe several 3D human poses even in case of strong occlusions or crops. Then, joint grouping and human pose estimation for an arbitrary number of people are performed using the associative embedding method. Our approach significantly outperforms the state of the art on the challenging CMU Panoptic. Furthermore, it leads to good results on the complex and synthetic images from the newly proposed JTA Dataset.

Keywords: multi-person, 3D, human pose, deep learning

1. Introduction

3D human pose estimation based on RGB images is a challenging task from the computer vision perspective. Recent Convolution Neural Network (CNN) based approaches [1, 2] achieve excellent performance in 2D human pose estimation thanks to large scale in the wild datasets. Nevertheless, methods for 3D human pose estimation require 3D ground truth that is only available using Motion Capture (Mocap) systems. Therefore, these methods have good performance in controlled environment but bad generalisation to real in the wild images. Furthermore, most of the 3D pose estimation methods are restricted to a single fully visible subject. In real-world scenarios, multiple people interact in cluttered or even crowded scenes containing both self-occlusions of the body and strong inter-person occlusions. Therefore, inferring the 3D pose of all the

subjects (without knowing in advance their number) from a single and monocular RGB image is a harder problem and recent single-person 3D human pose estimation methods fail in this case.

A natural approach is the decomposition of the multi-person ill-posed problem into multiple single-person 3D estimations. These top-down approaches are based on the generation of multiple pose proposals that are evaluated and refined in a second time [3]. Thus, they perform many redundant estimations and scale badly for a large number of subjects.

Another way to solve this problem is bottom-up approaches [4], [5], [6] that manage the whole scene in a single forward pass to give multi-person 3D human pose estimates. By their principle, they are more effective in managing occlusions between people and take advantage of context-related information to predict the different poses.

In the present article, we propose a new bottom-up approach that manages the whole scene in a single forward pass to give multi-person 3D human pose estimates. Our method is based on the Stacked Hourglass architecture [7] that has demonstrated its effectiveness for 2D human pose estimation. Single shot multi-person 3D human pose estimation is challenging as it needs to properly locate human joints and to regroup these estimations into final 3D skeletons. By associating the Hourglass architecture with a powerful joints grouping method named the associative embedding [2] and a robust multi-person 3D pose description [8], we design an end-to-end architecture that jointly performs 2D human joints detection, joints grouping and full body 3D human pose estimation even when the subjects are partially occluded or truncated by the image boundary. The proposed method surpasses state of the art results on the CMU-Panoptic [9] dataset and shows good results on the Joint Track Auto dataset[10], a synthetic but realistic dataset with a large number of people, various camera viewpoints and backgrounds.

2. Related Work

Human pose estimation is more and more studied as it is very useful for many applications (e.g. motion capture, human image synthesis, activity recognition, sign language recognition, robotics vision, etc.). In this section, we present recent deep learning approaches for 2D human pose estimation and single/multi-person 3D human pose estimation.

2D human pose estimation: Most methods for 2D human pose estimation extract probabilistic maps called heatmaps that estimate the probability of each pixel to contain a particular joint. At inference time, the 2D joint positions correspond to the local maxima of the heatmaps. Most of these methods [7, 11] are also iterative. A refined estimate of the heatmap is obtained from the previous estimates and the convolutional features. Wei *et al.* [11] refine the predictions over successive stages with intermediate supervision at each stage. The Stacked Hourglass networks [7] perform repeated bottom-up top-down processing with intermediate supervision.

Both top-down and bottom-up human approaches have been proposed for multi-person 2D human pose estimation. Top down methods [12, 13] first detect human bounding boxes and then estimate 2D human poses. Nevertheless, these methods fail when the detector fails, in particular when there are strong occlusions. Bottom-up approaches [1, 2] first estimate the 2D location of each joint and then associate them into full skeletons. Cao et al. [1] regress affinity between joints that means the direction of the bones in the image. Unlike this approach that needs complex post-processing to group joints, Newell et al. [2] propose to learn this association in an end-to-end network thanks to the Associative Embeddings.

Single-person 3D human pose estimation: Motivated by the recent advances in 2D human pose estimation, some existing approaches [14, 15, 16, 17, 18, 19, 20, 21, 22] use only 2D human poses estimated by other methods [7, 1] to predict 3D human poses. Chen and Ramanan [20] performs a nearest neighbour search on a given 3D pose library with a large number of 2D projections. Moreno-Noguer [21] formulate the problem of the 3D human pose estimation as a 2D to 3D distance matrix regression. Nie *et al.* [22] predict depth on joints using LSTM. Martinez *et al.* [14] lift 2D joints to 3D space using a deep residual neural network. Nevertheless, these approaches are limited by the 2D pose estimator performance and do not take into account important images clues, such as contextual information, to make the prediction.

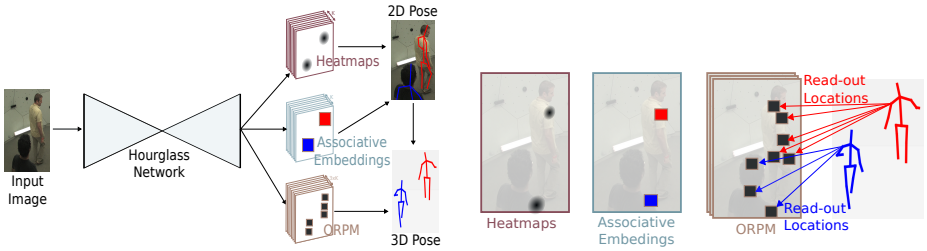
Other methods predict 3D human poses from images features[23, 24, 25, 26, 27]. Recent methods make this prediction directly from monocular images [28, 29, 30, 31, 32, 33] or from sequences of images [34, 35] using Convolutional Neural Networks. The learning procedure needs images annotated with 3D ground-truth pose. Since no large scale 3D in the wild annotated dataset exists, current approaches tend to overfit on the constrained environment they have been trained on. The existing in the wild approaches use either synthetic data [31, 32, 36] or are trained on both 3D and in the wild 2D datasets [8, 37, 38, 39, 40, 41, 42, 43]. Mehta *et al.* [8] use a pretrained 2D pose network to initialize the 3D pose regression network. Zhou *et al.* use geometric constraints [41] in a weakly supervised setting. Pavlakos *et al.* [42] take another approach by relying on weak 3D supervision in form of a relative 3D ordering of joints which can be easily annotated even for in the wild images. Yang *et al.* [44] use an adversarial loss that transfers the 3D human pose structures learned from the indoor annotated dataset to the in-the-wild images. Although performing well with a single fully visible subject, these methods fail with several interacting people that are at different image scale and that occult each other.

Multi-person 3D human pose estimation: In a top-down approach, Rogez *et al.* [3] generate human pose proposals that are further refined using a regressor. This approach performs many redundant estimations that need to be fused and scales badly for a large number of people. Zanfir *et al.* [5] estimate the 3D human shape from sequences of frames using a pipeline process followed by a 3D pose refinement based on a non-linear optimisation process and semantic constraints. MubyNet [4] is a bottom-up multi-task network that identifies joints

and learns to score their possible associations as limbs. These scores are used to solve a global optimisation problem that groups the joints into full skeletons following the human kinematic tree. Mehta *et al.* [6] propose an approach that predicts 2D heatmaps, part affinity fields [1] and Occlusions Robust Pose Maps (ORPM). This approach manages multi-person 3D human pose estimation even for occluded and cropped people. Nevertheless, the architecture used in [6] is not a stacked architecture while the stacking strategy [1, 2] performs well in the 2D context.

The proposed method deals with multi-person 3D human pose estimation. Unlike [5], it does not need sequence of images to refine the pose estimates. It is based on the stacked hourglass networks [7] devoted to mono-person 2D pose estimation and showing very good performance on this task. Thus, we extend this approach using the multi-person 3D poses description robust to occlusions proposed in [6] and the associative embedding [2] that groups joints in skeletons in a more effective way that part affinity fields [1] proposed in a 2D context. The final network architecture is notably trained in an end-to-end manner and the inference requires a single forward pass.

3. Proposed Method



(a) Model overview (b) Outputted maps for shoulder joint

Figure 1: The proposed model estimates full 3D skeletons for an arbitrary number of people. It predicts, for each joint, a 2D localisation map(heatmap), an associative embedding map and 3 ORPM. The associative embeddings maps contain different embedding values for joints belonging to different subjects. The ORPM store the 3D joints coordinates at different 2D locations. Best viewed in color.

3.1. Description

Given a monocular RGB image $\mathbf{I}$ of size W $\times$ H, we seek to estimate the 3D human poses $\mathcal{P}_I = \{P_i \,|\, i \in [1, \ldots, N]\}$ where N is the number of visible people, $P_i \in \mathbb{R}^{3\times K}$ are the 3D joints locations and K is the number of predicted joints. The 3D joint coordinates are expressed relatively to their parents joints in the kinematic tree and converted to pelvis relative locations for evaluation in a 3D coordinate reference oriented like the camera one. The model is composed of several stacked hourglass networks. The image is first sub-sampled to images

features **I'** of size $W' \times H'$ by convolution and pooling layers. Each hourglass module outputs heatmaps for 2D joints detection, ORPM for 3D joints localisation and associative embeddings maps for joint grouping, each map being of size $W' \times H'$. Except for the first hourglass that takes as input only image features, other hourglasses takes as input images features and the prediction of the previous hourglass that is refined. Fig. 1 depicts an overview of the proposed method.

3.2. Occlusions Robust Pose Maps

Suppose we have an image I and the corresponding 3D poses P_I. A good 3D pose representation to train a Convolutional Neural Network should have the following characteristics:

- a fixed dimension regardless of the number of people in the image;
- being robust to occlusions and crops.

To address these two problems, we adopt the ORPM formulation. For each joint, each hourglass network outputs three maps of dimensions $W' \times H'$, one for each X,Y,Z dimension. The size of these maps does not depend on the number of visible people which allows the estimation of the 3D pose of an arbitrary number of people. In these maps, the 3D joint coordinates of each person are stored at different 2D locations:

- at the 2D positions of the pelvis and the neck;
- at the 2D position of the joint;
- at the 2D positions of the joints belonging to the same limb.

For instance, the 3D coordinates of the wrist joints are stored in the wrist ORPM at the pelvis, the neck, the elbow and the shoulder 2D positions. This redundancy in the ORPM allows a robustness to occlusions and crops. Indeed, neck and pelvis are the best estimated and the less prone to occlusions.

At inference time, the 3D pose readout of a person is performed in two steps: a full 3D pose readout followed by a 3D pose refinement.

The full 3D pose readout is performed by reading the full person 3D pose at the following 2D positions in the ORPM:

- at the pelvis 2D position, if the pelvis is detected;
- at the neck 2D position, if the neck is detected and the pelvis is not.

If none of these two joints are detected, we take the mean skeleton in the training dataset as the full person 3D pose.

The full 3D pose readout is followed by the 3D pose refinement. During this step, for each joint, we refine the predicted 3D coordinates previously obtained by reading in the ORPM at one of the following 2D locations:

- at the joint 2D position in the ORPM if this 2D position is a valid readout location;
- at the 2D position of a joint belonging to the joint's limb. We take the extremity of the joint's limb and we go back in the kinematic tree until a valid readout location is found.

If no valid readout location is found in the joint's limb, the 3D coordinates are not refined. A 2D readout position is considered valid is it satisfies the following criteria:

- the confidence associated to the 2D predicted position of the joint is higher than a given threshold τ_C;
- the distance between the 2D joint position and the 2D position of the other joints must be less than a given distance τ_D;
- the 3D coordinates read at this 2D position in the ORPM must be anthropomorphically correct. In this purpose, we compute the mean length of each limb in the training dataset and we reject each predicted 3D coordinates that gives limbs whose length is too far from the corresponding computed mean.

3.3. Associative embedding

The network predicts for each joint a 2D heatmap and 3 ORMP for each X, Y, Z joint coordinates. This description is independent of the number of people. Now, we use the associative embedding to associate the joint to full skeletons. Predicted heatmaps contain peaks at the 2D joint positions of different subjects. To regroup the joints belonging to the same person, an additional output is added to the network for each joint corresponding to embeddings. Detections are then grouped by comparing the embedding values of different joints at each 2D peak position in the heatmap. If two joints have a close embedding value, they belong to the same person. The network is trained to perform this grouping by predicting close embeddings for joints belonging to the same person and distant embeddings for joints of distinct people.

Formally, let $E_k \in \mathbb{R}^{W' \times H'}$ be an embedding map predicted by the network for the k^{th} joint and $e_k(\boldsymbol{x})$ be the embedding value at the 2D position $\mathbf{x}$. Let us consider an image composed of N people, each having K joints. Let $\boldsymbol{x_{k,n}}$ be the 2D ground-truth position of the k^{th} joint of the person n. We refer by *reference embedding*, the predicted embedding of a person obtained as the mean of its embedding's joints:

$$\overline{e}_n = \frac{1}{K} \sum_k e_k(\boldsymbol{x_{k,n}}) \tag{1}$$

The grouping loss is then defined by:

$$\mathcal{L}_{AE} = \frac{1}{NK}\sum_{n}\sum_{k}(\overline{e}_n - e_k(x_{k,n}))^2 + \frac{1}{N^2}\sum_{n}\sum_{n'\neq n}\exp\left(-\frac{1}{2\sigma^2}(\overline{e}_n - \overline{e}_{n'})^2\right) \tag{2}$$

The first term of equation (2) corresponds to a pull loss that brings similar embeddings for joints belonging to a same person and the second part corresponds to a push loss that gives different embeddings to joints of different subjects. σ is a parameter giving more or less importance to the push loss.

3.4. Network loss

We learn jointly the three following tasks: i) 2D joint localisation by predicting heatmaps; ii) 3D joint coordinates estimation with ORPM prediction; iii) Joint grouping with associative embedding prediction. The network loss is then:

$$\mathcal{L}_{\text{3DMP}} = \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{ORPM}} + \lambda_{\text{AE}}\,\mathcal{L}_{\text{AE}} \tag{3}$$

Where $\mathcal{L}_{\text{2D}}$ is the euclidean distance between the ground-truth 2D heatmaps and the predicted 2D heatmaps, $\mathcal{L}_{\text{ORPM}}$ is the euclidean distance between the predicted ORMP and the ground-truth ORMP and $\mathcal{L}_{\text{AE}}$ is the loss defined by equation (2). And $\lambda_{\text{AE}} = 0.001$ is the weight of the Associative Embeddings loss.

3.5. Multi-Scale Inference

Although being single-shot and working well when there are a reduced number of people that are close to the camera, our method with a single scale inference tends to fail in complex and crowded images like those from the JTA dataset. In these images, visible people are at very different distances from the camera. Consequently, these people are projected with very different pixel resolutions and the model has difficulties to handle properly all these scales with a single image resolution. To handle these cases, Multi-Scale Heatmaps, Multi-Scale Associative Embeddings maps and Multi-Scale ORPM are computed.

Suppose that we have an input image $\mathbf{I}$ for which we want to extract Multi-Scale Heatmaps, Multi-Scale Associative Embeddings maps and Multi-Scale ORPM. Let $\mathbf{S} = s_1, s_2, \ldots, s_M$ the scale pyramid for which we want to compute these maps, s_M being the highest resolution scale.

First, for each scale s_i, we compute $\mathbf{HM_{s_i}} \in \mathbb{R}^{K\times W_{s_i}\times H_{s_i}}$, $\mathbf{A_{s_i}} \in \mathbb{R}^{K\times W_{s_i}\times Hs_i}$, $\mathbf{O_{s_i}} \in \mathbb{R}^{3\times K\times W_{s_i}\times Hs_i}$ respectively the predicted heatmaps, associative embedding maps and ORPM for scale s_i. Each $\mathbf{H_{s_i}}$, $\mathbf{A_{s_i}}$ and $\mathbf{O_{s_i}}$ is resized to maps $\mathbf{HM'_{s_i}}$, $\mathbf{A'_{s_i}}$ and $\mathbf{O'_{s_i}}$ that match the resolution of scale s_M.

The Multi-Scale Heatmaps are the mean of the rescaled heatmaps. Let $\mathbf{MSH} \in \mathbb{R}^{K\times W_{s_M}\times H_{s_M}}$ be the Multi-Scale Heatmaps, $msh(j, \boldsymbol{x})$ be the value of $\mathbf{MSH}$ at position $\boldsymbol{x}$ for joint j and $h'_{s_i}(j, \boldsymbol{x})$ be this value for the rescaled heatmaps $\mathbf{HM'_{s_i}}$. Then, we have :

$$msh(j, \boldsymbol{x}) = \frac{1}{M}\sum_{i=1}^{M} h'_{s_i}(j, \boldsymbol{x}) \tag{4}$$

The Multi-Sacle Associative Embeddings maps are the concatenation of the rescaled associative emedings maps $\mathbf{A'_{s_i}}$.

In order to compute the Multi-Scale ORMP $\mathbf{MSO} \in \mathbb{R}^{3\times K\times W_{s_M}\times H_{s_M}}$, we cannot compute a simple average like done for the heatmaps. Indeed, if a person is detected at a given scale but not in another one, if we simply compute the average between the ORPM at each scale, the well estimated 3D pose at one scale could be altered by this operation. To avoid this, the mean is weighted by the predicted heatmaps and we take into account the different readout locations induced by the ORPM formulation. This way, more the model is confident about a predicted joint at a given scale, more the ORPM at this scale will contribute to the **MSO**. Let $mso(c, j, \boldsymbol{x})$ be the value of **MSO** for coordinate c (X, Y or Z coordinate) and joint j at 2D position $\boldsymbol{x}$, $O'_{s_i}(c, j, \boldsymbol{x})$ be this value for the rescaled ORPM $\mathbf{O'_{s_i}}$ and $RL^j = rl^j_1, rl^j_2, \ldots, rl^j_{L_j}$ be the set of readout locations for the joint j in the ORPM. Then, we have:

$$mso(c, j, \boldsymbol{x}) = \frac{\sum_{i=1}^{M}\sum_{l=1}^{L_j} h'_{s_i}(rl^j_l, \boldsymbol{x}) o'_{s_i}(c, j, \boldsymbol{x})}{\sum_{i=1}^{M}\sum_{l=1}^{L_j} h'_{s_i}(rl^j_l, \boldsymbol{x})} \tag{5}$$

3.6. Final prediction

Once the network is trained, the final prediction is obtained in several stages. First, a non-maximum suppression is applied on the heatmaps to obtain the set of joint detections. Then, all the neck embeddings are read from the neck embedding map at the predicted neck 2D positions. This pool of 2D neck positions with their corresponding embedding gives the initial set of detected people. The other joints associated to these necks need now to be found. Each person is characterised by its reference embedding. The next joint associated to a given person is the one having the highest detection score and having a distance with the person embedding lower than a given threshold τ_{AE}. We repeat this step until there is no more joint that respects these two criteria. Once this process is done, the non-associated joints are used to create a new pool of people. At the end, the 2D pose of each person is obtained and used to read the 3D pose in the ORPM as described in Section 3.2.

4. Experiments

In this paper, we address the problem of single shot multi-person 3D human pose estimation. To evaluate our method, we perform separate experiments on:

- multi-person 3D human pose estimation in a controlled environment (CMU-Panoptic dataset [9]); some images are depicted in Figure 2.
- multi-person 3D human pose estimation in virtual environments with many people (JTA dataset)[10]. This dataset is more complex and richer than the previous one. Some images are shown in Figure 3. No previous method for 3D human pose estimation has been evaluated on this dataset to the best of our knowledge.

Evaluation Metrics: To evaluate our Multi-Person 3D pose approach, we use two metrics. The first one is the Mean per Joint Position Error (MPJPE) that corresponds to mean Euclidean distance between ground truth and prediction for all people and all joint. The second one is the 3D PCK which is 3D extension of the Percentage of Correct Keypoints (PCK) metric used for 2D Pose evaluation, as well. A joint is considered correctly estimated if the error in its estimation is less than 150mm. If an annotated subject is not detected by our approach, we consider all of its joints to be incorrect in the 3D PCK metric.

Training Procedure: The method was implemented with PyTorch. The hourglass component is based on the public code in [2]. We used four stacked hourglasses in our model, each one outputting 2D heatmaps, ORPM and associative embeddings. We trained the model using mini-batches of size 30 on 8 Nvidia Titan X GPU during 240k iterations. We used the Adam[45] optimiser with an initial learning rate of 10^{-4}.

4.1. Multi-person 3D pose estimation on CMU-Panoptic

CMU Panoptic [9] is a dataset containing images with several people performing different scenarios (playing an instrument, dancing, etc.) in a dome where several cameras are placed. This dataset is challenging because of complex interactions and difficult camera viewpoints. We evaluate our model following these protocols:

- Panoptic-1 protocol: it is the protocol used in [5, 4]. The model is evaluated on 9600 frames from HD cameras 16 and 30 and for 4 scenarios: Haggling, Mafia, Ultimatum, Pizza. The model is trained on the other 28 HD cameras of this dataset.

- Panoptic-2 protocol: This protocol is an extension of the previous one. Instead of evaluating on a subset of arbitrary selected frames, we evaluate on the entire sequences from cameras 16 and 30. The training dataset in this protocol is the frames from all the HD cameras (except cameras 16 and 30) for the Haggling, Mafia, Ultimatum, Pizza scenarios. The model is evaluated on the same scenarios by taking one frame every ten frames from HD cameras 16 and 30.

- Panoptic-3 protocol: Previous protocols use a large number of training cameras. To evaluate the robustness to the number of cameras and to the amount of training data, we propose protocol Panoptic-3. The model is trained on the Haggling, Mafia, Ultimatum, Pizza scenarios but only a subset of the training cameras is used:

  - Panoptic 3a: HD cameras 0, 2, 4, 6, 8, 10, 12, 14, 18, 20, 22, 24, 26 and 28 are used during training
  - Panoptic 3b: HD cameras 0,4,8,12,20,24 and 28 are used during training
  - Panoptic 3c: HD cameras 0,8, and 24 are used during training

Method	Haggling	Mafia	Ultimatum	Pizza	Mean
[30]	217.9	187.3	193.6	221.3	203.4
[5]	140.0	165.9	150.7	156.0	153.4
[4]	72.4	78.8	66.8	94.3	72.1
Ours, full	**70.1**	**66.6**	**55.6**	**78.4**	**68.5**

Table 1: Mean per joint position error (MPJPE) in mm on the Panoptic Dataset following Panotic-1 Protocol

The test set is the same as Panoptic 2.

- Panoptic-4 protocol : In the previous protocols, the model is trained and evaluated on the same scenarios. To evaluate the robustness to an unseen scenario in new camera viewpoints, we propose the Panoptic-4 protocol. The training dataset in this protocol is the frames from all the HD cameras (except cameras 16 and 30) from the Haggling, Mafia and Ultimatum scenarios. The model is evaluated on the pizza scenario by taking one frame every ten frames from HD cameras 16 and 30.

Comparison with prior work: On Panoptic-1 protocol, our model improves the results over the recent state of the art methods on all the scenarios (Table 1). It shows a global improvement of 5.0% compared to [4]. Note that unlike [5] we do not learn on any frame from the cameras 16 and 30 and on any external data. Actually, the proposed model does not need a trained attention readout process thanks to the effective ORPM readout process.

Ablative studies: Table 2 provides ablative results of our method following Panoptic-1 protocol on the Haggling, Mafia, Ultimatum and Pizza scenarios. Firstly, we present the results obtained by stacking one, two or three hourglass modules. Each time an hourglass module is added, the Mean per Joint Position Error (MPJPE) decreases (from 91.8 mm for one hourglass module to 68.5 mm for our full four hourglass modules model). This shows the importance of the stacking scheme and the refinement process in the model architecture. The penultimate line of this table shows the results obtained with four hourglass modules and a Naive Readout (NR) in the ORPM, that means when the 3D joint coordinates are read directly from their 2D positions. Because of frequent crops and occlusions in the panoptic dataset, this model has poor performance with an MPJPE of 118.8 mm. This proves the importance of the ORPM storage redundancy to manage occlusion. Our complete model(last row) with four hourglass modules and the readout procedure described in Section 3.2 has the lowest MPJPE (68.5mm)

Examples of 3D human pose estimations on the Panoptic dataset are shown in Figure 2. Our method can estimate the 3D pose of multiple people even in case of truncation (1st, 2nd and last rows) or people overlap (2nd and 4th rows)

Robustness to the number of training cameras : Protocols Panoptic 1 and 2 results are obtained by using a large number of training cameras. What

Method	Nb of HG	ORPM	Haggling	Mafia	Ultimatum	Pizza	Mean
Ours, 1-HG	1	✓	92.3	86.1	82.7	103.8	91.8
Ours, 2-HG	2	✓	77.1	74.8	68.0	89.8	78.3
Ours, 3-HG	3	✓	72.4	72.4	60.12	85.2	73.8
Ours, NR	4	×	101.5	124.2	105.7	130.3	118.8
Ours, full	4	✓	**70.1**	**66.6**	**55.6**	**78.4**	**68.5**

Table 2: Mean per joint position error (MPJPE) in mm on the Panoptic Dataset following Panoptic-1 protocol. (i-HG stands for i stacked hourglasses and NR for Naive Readout).

Protocol	Haggling	Mafia	Ultimatum	Pizza	Mean
Panoptic 2	78.3	60.7	84.2	78.3	68.1
Panoptic 3a	82.4	64.3	88.7	82.2	72.3
Panoptic 3b	84.0	74.2	87.4	92.0	76.4
Panoptic 3c	149.4	151.3	155.5	167.9	150.9
Panoptic 4				79.4	79.4

Table 3: Mean per joint position error (MPJPE) in mm on the Panoptic Dataset following Panoptic-2, Panoptic-3 and Panoptic-4 protocols

is the robustness of our model when using a reduced number of cameras ? Table 3 provides Panoptic 3 protocol results. Panoptic 3a and 3b results show that even by using only half and fourth of the training cameras, the MPJPE is only increased respectively by 6.9% and 12.2%. On the other hand, where only 3 training cameras are used, the MPJPE is 2.2 times greater than the Panoptic 2 MPJPE. This number of cameras is insufficient to learn such a complex task. Even single person 3D human pose models are trained on datasets[46, 8] that provides images from four cameras or more.

Performance on an useen scenario: Protocols Panoptic 1,2 and 3 show the ability of the model to generalise to unseen camera viewpoints. Panoptic 4 results show the ability of the model to generalise to new scenarios. The model is trained only on the Haggling, Mafia and Ultimatum scenarios and evaluated on the unseen Pizza scenario. The Panoptic 4 MPJPE (79.4) is close the MPJPE obtained on the Panoptic 2 protocol for the Pizza scenario showing that model does not overfeat on the training scenarios and can generalise to new ones.

4.2. Multi-person 3D pose estimation on JTA dataset

JTA (Joint Track Auto) is a dataset for human pose estimation and tracking in urban environment. It was collected from the realistic video-game the Grand Theft Auto V and contains 512 HD videos of 30 seconds recorded at 30 fps. The collected videos feature a vast number of different body poses, in several urban scenarios at varying illumination conditions and viewpoints. People perform different actions like walking, sitting, running, chatting, talking on the phone, drinking or smoking. Each image contains a number of people ranging between 0 and 60 with an average of more than 21 people. The distance from the camera ranges between 0.1 to 100 meters, resulting in pedestrian heights between 20

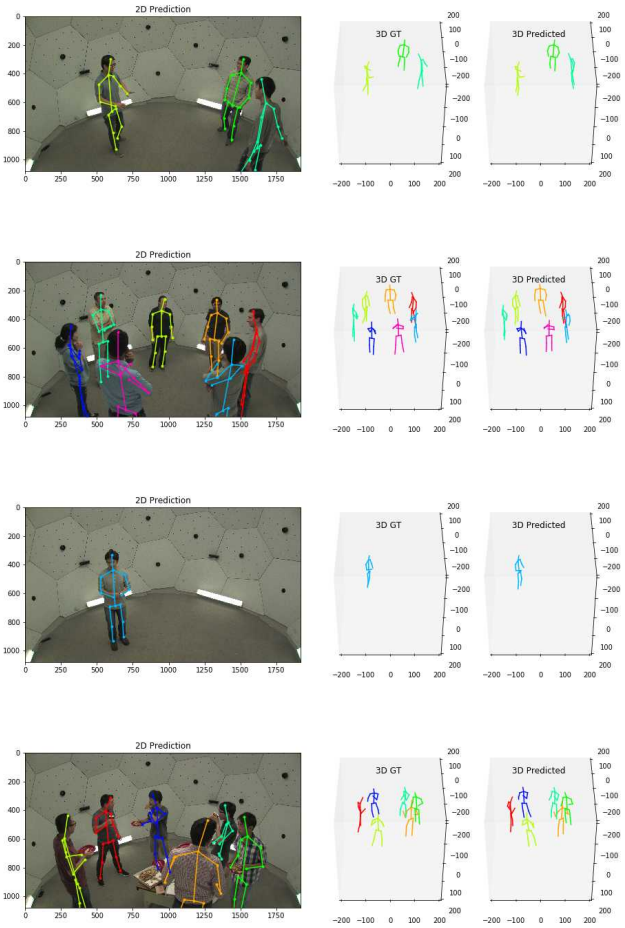


Figure 2: Multi-person poses predicted by our approach on the CMU-Panoptic Dataset. Ground truth translation and scale are used for visualisation. The first column corresponds to the input image with the predicted 2D pose. The second column corresponds to the ground truth 3D poses and the last column to the predicted 3D poses. These examples show that our approach works with a variable number of people in the image and can predict the 3D coordinates of joints that are not visible in the image thanks to the ORPM redundancy. Best viewed in color.

and 1100 pixels. None existing (virtual or real) dataset with annotated 3D pose is comparable with JTA dataset in terms of number of people per image, people and background variability. As far as we know, we are the first to demonstrate the ability of a trained model to deal with such complex and rich environments with many people at different camera distances and with different resolutions. 256 videos are used for training and 128 for testing (the remaining 128 videos are used for validation). From the testing videos, we take one frame every ten frames for the evaluation.

Table 4 presents per camera distance results on the the JTA Dataset. We evaluate our model on this dataset at different resolutions (S1=512px, S2=1024px and S3=1536px) and also with the multi-scale inference described in section 3.5. The images from this dataset contain a large number of people in various distances from the camera. The distance from the camera can have a significant impact on the performance of a 3D human pose estimator. Indeed, distant people require higher image resolution and are more likely to be occulted. For this reason, we provide in Table 4 results for people in different ranges of distance from the camera. Note that our testing set contains 262510 people. Among these people, 10% have a distance from the camera less than 10 meters, 23% have a distance from the camera between 10 and 20 meters, 21% have a distance from the camera between 20 and 30 meters, 14% have a distance from the camera between 30 and 40 meters and 31% have a distance from the camera greater than 40 meters.

The resolution having the best overall 3D PCK is the resolution S2 with a 3D PCK of 37.8%. This resolution performs a good compromise to estimate the pose of the high resolution people that S3 cannot handle properly and low resolution people that are too small from scale S1. Resolution S1 has the best results for people that are close to the camera (less than 10 meters) with an MPJPE of 165.2mm and a 3D PCK of 68.5%. Resolution S2 has the best results for people that have a distance from the camera between 10 and 20 meters with a 3D PCK of 62.3% and an MPJPE of 194.50. Resolution S3 has the best results for people that are far from the cameras (greater than 20 meters). These results show that each resolution is adequate to a given range of people distance and consequently to a resolution of people.

The multi-scale inference (MSI) improves the overall 3D PCK and MPJPE. The 3D PCK goes from 37.8 to 43.9 for the MSI and the MPJPE goes from 258.9mm to 193.5mm. MSI has better results than scale S2 and S3 for close to the camera people (less than 10 meters) taking advantages from poses estimated from scale S1 but without improving over this scale for these people. MSI has equivalent results to scales S1 and S2 for people that have a distance from the camera between 10 and 20 meters and significantly better than scale S3 for these people. It has worse results than scales S2 and S3 for people that have a distance from the camera between 20 and 40 meters but it surpasses all the scales for people that have a distance from the camera greater than 40 meters.

Joint-wise analysis (Table 5) shows that the results are unequal from one joint to another one. Regardless of the distance to the camera, spines and hips are always the best estimated joints. These articulations have a reduced

Scale	Distance to camera	MPJPE	3D PCK
S1 (512 px)	<10m	165.2	68.5
	>10m and <20m	220.6	61.6
	>20m and <30m	358.7	42.2
	>30m and <40m	409.7	36.0
	>40m	382.1	32.2
	All	**294.0**	**33.1**
S2(1024px)	<10 m	275.53	43.5
	>10m and <20m	194.50	62.3
	>20m and <30m	281.5	51.25
	>30m and <40m	358.8	41.0
	>40m	368.2	35.5
	All	**258.9**	**37.8**
S3(1536px)	<10m	319.0	33.9
	>10m and <20m	231.16	49.4
	>20m and <30m	222.75	53.3
	>30m and <40m	269.1	47.5
	>40m	305.90	38.8
	All	**274.3**	**34.8**
Multi Scale Inference(MSI)	<10m	175.5	55.8
	>10m and <20m	220.6	61.6
	>20m and <30m	358.6	42.2
	>30m and <40m	409.7	36.0
	>40m	262.12	41.7
	All	**193.5**	**43.9**

Table 4: MPJPE and 3D PCK on the JTA dataset. Results are provided per scale and per camera distance that means by taking into account in the metrics computation only the people that are in the corresponding distance range from the camera.

variability compared to the extremity joints like wrists and ankles that have the worst MPJPE and 3D PCK. Indeed, since the 3D joint coordinates are expressed relatively to their parents joints in the kinematic tree and converted to pelvis relative locations, errors in the estimation of a parent joint impact the estimation of all its descendent in the kinematic tree.

Examples of 3D human pose estimations on the JTA dataset are shown in Figure 3. Our method can estimate the 3D pose in several urban scenarios at varying illumination conditions and viewpoints. Nevertheless, very far people are not detected and the method fails in case of crowded people.

Figure 4 shows qualitative results on natural images. The trained model is able to predict 3D human poses even for real in the wild images.

5. Conclusion

We have presented a single shot trainable model for multi-person 3D human pose estimation in real environment with various camera viewpoint conditions, strong occlusions and several social activities or in virtual but very realistic environment with a vast number of body poses, and several urban scenarios at varying illumination conditions and viewpoints. 2D and 3D human joints are predicted using heatmaps and ORPM which have proven their ability to manage

Figure 3: Qualitative results (reprojected 3D poses) of our approach shown on the test set of JTA Dataset. Our 3D estimations are relative to the pelvis. Ground truth translation and scale are used for visualisation. Our approach predict 3D poses in several urban scenarios at varying illumination conditions and viewpoints and for low resolution people. Nevertheless, very far people are not detected and the method fails in case of crowded people. Best viewed in color.

Distance to camera	Metric	head	neck	clavicles	shoulders	elbows	wrists	spines	hips	knees	ankles	all
>0	MPJPE	196.5	174.7	174.9	215.3	264.6	329.4	42.3	76.3	253.2	425.5	193.5
	3D PCK	41.1	44.6	44.9	33.8	27.2	19.0	74.4	73.9	25.7	8.9	43.9
<10m	MPJPE	131.7	195.1	191.8	219.5	218.7	254.6	45.8	66.97	236.1	395.9	175.5
	3D PCK	68.1	48.1	48.5	37.5	39.5	30.6	94.2	94.0	29.0	7.3	55.8
>10m and <20m	MPJPE	231.2	188.6	192.5	239.7	297.14	365.7	47.1	86.0	297.2	505.3	220.6
	3D PCK	60.0	66.5	65.1	52.8	45.5	33.5	91.9	87.8	45.2	24.9	61.6
>20m and <30m	MPJPE	392.1	302.3	309.2	385.9	484.2	581.4	73.0	142.0	489.5	827.3	358.6
	3D PCK	36.1	45.6	42.8	23.9	16.6	9.9	85.0	73.8	18.0	7.9	42.2
>30m and <40m	MPJPE	451.1	345.9	352.1	443.3	552.0	650.7	84.21	166.4	561.9	945.8	409.7
	3D PCK	28.7	37.8	34.4	14.3	8.8	4.7	82.1	65.9	10.3	4.6	36.0
>40m	MPJPE	248.0	200.3	212.1	310.8	410.6	505.2	49.7	119.0	324.7	528.5	262.12
	3D PCK	39.2	50.1	45.3	18.0	11.1	6.0	89.7	72.1	13.1	4.5	41.7

Table 5: Joint wise MPJPE and 3D PCK on the JTA Dataset of our approach with the Multi-Scale Inference

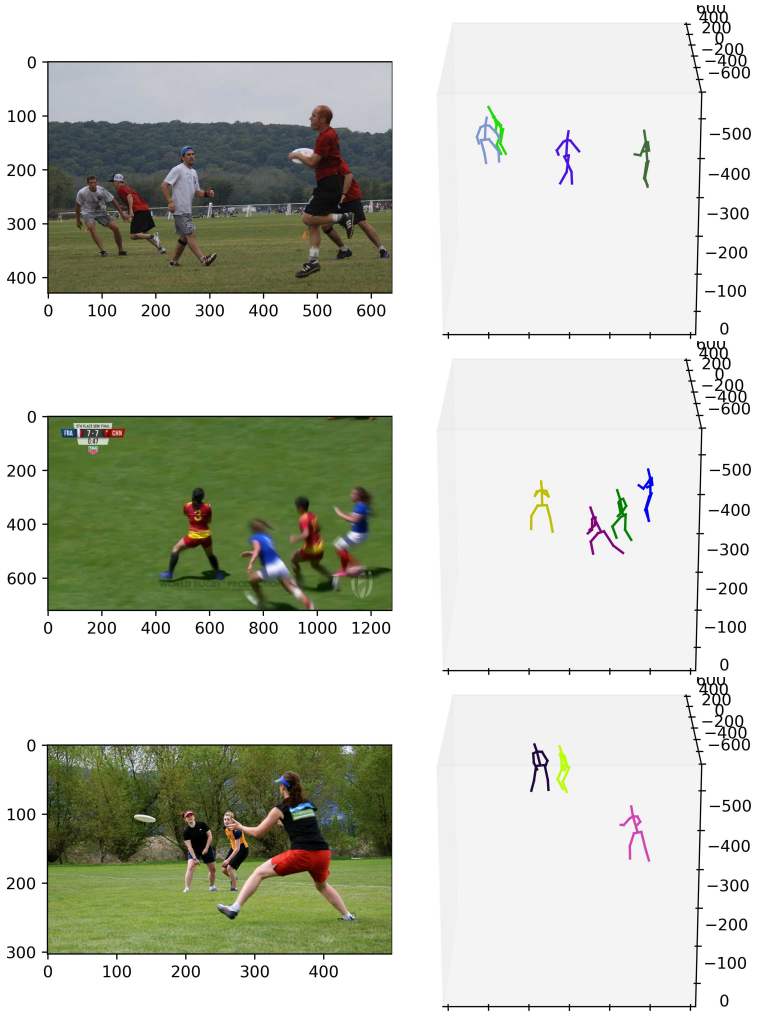


Figure 4: Multi-person poses predicted by our approach on natural images. The first column corresponds to the input image. The second column corresponds to predicted 3D poses. Best viewed in color.

occlusions. The difficult problem of associating joints to people skeletons is managed using the recent associative embeddings method. The same stacked network jointly learns and estimates, in an end-to-end manner, 2D human poses and 3D human poses exploiting the complementarity of these tasks.

The experiments provided in this work have proven the importance of the stacking scheme and the ORMP formulation, validating the proposed network architecture. Furthermore, large-scale experiments, on the CMU Panoptic dataset, demonstrate that the proposed approach results surpass those of the state of the art. Nevertheless, the experiments on the JTA Dataset, although being correct for high resolution people close to the camera, show that complex urban scenarios with many people at different image resolution remains a challenge for our approach. Thus, we are working on an extension of this method to solve these difficult cases.

References

[1] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, CVPR (2017).

[2] A. Newell, Z. Huang, J. Deng, Associative embedding: End-to-end learning for joint detection and grouping, NIPS (2017).

[3] G. Rogez, P. Weinzaepfel, C. Schmid, Lcr-net: Localization-classification-regression for human pose, CVPR (2017).

[4] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, C. Sminchisescu, Deep network for the integrated 3d sensing of multiple people in natural images, NIPS (2018).

[5] A. Zanfir, E. Marinoiu, C. Sminchisescu, Monocular 3d pose and shape estimation of multiple people in natural scenes–the importance of multiple scene constraints, CVPR (2018).

[6] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, C. Theobalt, Single-shot multi-person 3d body pose estimation from monocular rgb input, 3DV (2017).

[7] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, ECCV (2016).

[8] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3d human pose estimation in the wild using improved cnn supervision, 3DV (2017).

[9] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, et al., Panoptic studio: A massively multiview system for social interaction capture, PAMI (2019).

[10] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, R. Cucchiara, Learning to detect and track visible and occluded body joints in a virtual world, ECCV (2018).

[11] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, CVPR (2016).

[12] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, Towards accurate multi-person pose estimation in the wild, CVPR (2017).

[13] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, ICCV (2017).

[14] J. Martinez, R. Hossain, J. Romero, J. J. Little, A simple yet effective baseline for 3d human pose estimation, ICCV (2017).

[15] H.-S. Fang, Y. Xu, W. Wang, X. Liu, S.-C. Zhu, Learning pose grammar to encode human body configuration for 3d pose estimation, AAAI Conference on Artificial Intelligence (2018).

[16] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M. J. Black, Keep it smpl: Automatic estimation of 3d human pose and shape from a single image, ECCV (2016).

[17] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, F. Moreno-Noguer, Single image 3d human pose estimation from noisy observations, CVPR (2012).

[18] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, W. Gao, Robust estimation of 3d human poses from a single image, CVPR (2014).

[19] V. Ramakrishna, T. Kanade, Y. Sheikh, Reconstructing 3d human pose from 2d image landmarks, ECCV (2012).

[20] C.-H. Chen, D. Ramanan, 3d human pose estimation= 2d pose estimation+ matching, CVPR (2017).

[21] F. Moreno-Noguer, 3d human pose estimation from a single image via distance matrix regression, CVPR (2017).

[22] B. X. Nie, P. Wei, S.-C. Zhu, Monocular 3d human pose estimation by predicting depth on joints, ICCV (2017).

[23] A. Agarwal, B. Triggs, 3d human pose from silhouettes by relevance vector regression, CVPR (2004).

[24] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, P. H. Torr, Randomized trees for human pose detection, CVPR (2008).

[25] C. Sminchisescu, A. Jepson, Generative modeling for continuous non-linearly embedded visual inference, ICML (2004).

[26] L. Bo, C. Sminchisescu, A. Kanaujia, D. Metaxas, Fast algorithms for large scale conditional 3d prediction, CVPR (2008).

[27] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter sensitive hashing, ICCV (2003).

[28] G. Pavlakos, X. Zhou, K. G. Derpanis, K. Daniilidis, Coarse-to-fine volumetric prediction for single-image 3d human pose, CVPR (2017).

[29] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, C. Theobalt, Vnect: Real-time 3d human pose estimation with a single rgb camera, TOG (2017).

[30] A.-I. Popa, M. Zanfir, C. Sminchisescu, Deep multitask architecture for integrated 2d and 3d human sensing, CVPR (2017).

[31] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, B. Chen, Synthesizing training images for boosting human 3d pose estimation, 3DV (2016).

[32] G. Rogez, C. Schmid, Mocap-guided data augmentation for 3d pose estimation in the wild, NIPS (2016).

[33] S. Li, A. B. Chan, 3d human pose estimation from monocular images with deep convolutional neural network, ACCV (2014).

[34] B. Tekin, A. Rozantsev, V. Lepetit, P. Fua, Direct prediction of 3d body poses from motion compensated sequences, CVPR (2016).

[35] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, K. Daniilidis, Sparseness meets deepness: 3d human pose estimation from monocular video, CVPR (2016).

[36] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, C. Schmid, Learning from synthetic humans, CVPR (2017).

[37] X. Sun, J. Shang, S. Liang, Y. Wei, Compositional human pose regression, ICCV (2017).

[38] E. Simo-Serra, A. Quattoni, C. Torras, F. Moreno-Noguer, A joint model for 2d and 3d pose estimation from a single image, CVPR (2013).

[39] F. Zhou, F. De la Torre, Spatio-temporal matching for human detection in video, European Conference on Computer Vision (2014).

[40] B. Tekin, P. Márquez-Neila, M. Salzmann, P. Fua, Learning to fuse 2d and 3d image cues for monocular body pose estimation, ICCV (2017).

[41] X. Zhou, Q. Huang, X. Sun, X. Xue, Y. Wei, Towards 3d human pose estimation in the wild: a weakly-supervised approach, ICCV (2017).

[42] G. Pavlakos, X. Zhou, K. Daniilidis, Ordinal depth supervision for 3d human pose estimation, CVPR (2018).

[43] Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial learning of structure-aware fully convolutional networks for landmark localization (2017).

[44] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, X. Wang, 3d human pose estimation in the wild by adversarial learning, CVPR (2018).

[45] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization (2014).

[46] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 1325–1339.