DEEP, ROBUST AND SINGLE SHOT 3D MULTI-PERSON HUMAN POSE ESTIMATION FROM MONOCULAR 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 pose estimation, from monocular RGB images. Our 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. Our approach leverages and extends the Stacked Hourglass Network and its multi-scale feature learning to manage multi-person situations. Thus, we exploit the Occlusions Robust Pose Maps (ORPM) to fully describe several 3D human poses even in case of strong occlusions or cropping. Then, joint grouping and human pose estimation for an arbitrary number of people are performed using associative embedding. We evaluate our method on the challenging CMU Panoptic dataset, and demonstrate that it achieves better results than the state of the art.

Index Terms— multi-person, 3D, human, pose, estimation.

1. INTRODUCTION

3D human pose estimation based on RGB images is a challenging task from the computer vision perspective. 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.

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. This allows to manage occlusions between people and to take advantage of context-related information to predict the different poses. Our method is based on the Stacked Hourglass architecture [1] 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 [3], we design an end-to-end architecture that jointly performs 2D human joints detection, joints grouping and 3D human pose estimation. We trained the model on a large scale dataset with real and complex human interactions and occlusions. The proposed method surpasses state of the art results on the CMU-Panoptic [4] dataset.

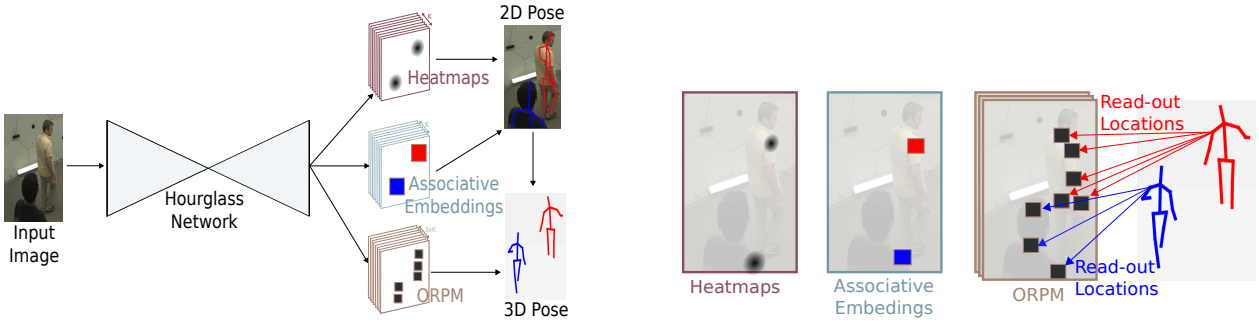
2. RELATED WORK

Human pose estimation is more and more studied as it is very useful for many applications (e.g., activity 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: Among the CNN based architectures proposed for single-person 2D human pose estimation [1, 5, 6], the Stacked Hourglass networks [1] are widely used as they process features across scales and efficiently capture spatial relationships between joints. Furthermore, its stacked architecture allows successive refinements of the pose estimates.

Both top-down and bottom-up human approaches have been proposed for multi-person 2D human pose estimation. Top down methods [7, 8] 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 [9, 2] first estimate the 2D location of each joint and then associate them into full skeletons. Unlike the part affinity fields [9] that need complex post-processing to group joint, Newell *et al.* [2] propose to learn this association in an end-to-end network thanks to the Associative Embedings.

Single-person 3D human pose estimation: Some existing approaches [10, 11] use only 2D human poses estimated by other methods [1, 9] to predict 3D human poses. These approaches do not take into account important images clues, such as contextual information, to make the prediction. Other methods [12, 13, 14] directly predict 3D human poses from



(a) Model overview

(b) Outputted maps for shoulder joint

Fig. 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.

images. 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 [15] or are trained on both the 3D and the in the wild 2D datasets [3, 16, 17, 18] by multi-modal 3d human pose supervision [3] or by using geometric constraints [16], ordinal depth supervision [17] or an adversarial loss [18].

Multi-person 3D human pose estimation: In a top-down approach, Rogez *et al.* [19] generates 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.* [20] 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 [21] 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.* [22] propose an approach that predicts 2D heatmaps, part affinity fields [9] 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 [22] is not a stacked architecture while the stacking strategy [9, 2] performs well in the 2D context.

The proposed method deals with multi-person 3D human pose estimation. Unlike [20], it does not need sequence of images to refine the pose estimates. It is based on the stacked hourglass networks [1] 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 [22] and the associative embedding [2] that groups joints in skeletons in a more effective way that part affinity fields [9] 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

3.1. Description

Given a monocular RGB image I of size W $\times$ H, we seek to estimate the 3D human poses $\mathcal{P} = \{P_i \,|\, i \in [1, \ldots, N]\}$ 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 frame oriented like the camera frame. 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 convolutions and pooling layers. Each hourglass module outputs heatmaps for 2D joints detection, ORPM for 3D joint 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, each hourglass 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

A 3D location map allows to store 3D coordinates of a joint in its 2D corresponding position. For each joint, three location maps (one for each coordinate X, Y, Z) as well as 2D heatmaps encoding 2D position in the form of confidence maps are predicted. In a basic 3D location map like the one used in [13], the 3D joint position is obtained in its 2D corresponding position in the map. However, this formulation sup-

poses that all the articulations are visible and is not adapted in case of strong occlusions. Furthermore, if a person is cropped and some of its joints are not visible, it is then impossible to predict their 3D positions even if they can be deduced from the global person posture. To manage these cases, the Occlusions Robust Pose Maps (ORPM) add redundancy and use the concept of valid read-out location. A 2D readout position of a joint is considered valid if the joint is not cropped, if the confidence score associated to the 2D predicted position of this joint is high, and if this joint is not too close to another joint in the image. Thus, the 3D location of a joint can be read in the following 2D positions of the corresponding ORPM:

- At the 2D predicted position itself. For instance, the elbow 3D coordinates are read from the elbow 2D position in the ORPM.
- At the 2D position of another joint of the limb if the 2D joint position is not a valid readout location. We start from the extremity of the limb and we go back in the kinematic tree until we found a valid readout position in the limb. For example, the wrist coordinates can be read at the 2D positions of the elbow or the shoulder.
- At the neck or the pelvis predicted 2D positions if the joints belonging to the limb are also not valid read-out locations. Thus, the 3D person skeleton can be obtained entirely at the 2D predicted positions of the neck and the pelvis. These two joints have been chosen because they are the most easily detected and the less prone to occlusions.

If the 2D positions of the pelvis and the neck are not valid read-out locations and if the person is detected in the 2D heatmaps, the model predicts the mean 3D position of the joint in the training dataset.

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 $e_k(x)$ be the embedding value at the 2D position x. Let us consider an image composed of N people, each having K joints. Let $x_{k,n}$ 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:

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

The grouping loss is then defined by:

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

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:

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

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 the λ's are the weights of the respective sub-losses.

3.5. 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. We repeat this step until there is no more joint that respects this 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

Datasets: We provide quantitative results on the CMU Panoptic [4] and Human 3.6M [23] datasets. CMU Panoptic [4] 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 the same protocol than [20, 21], that is on 9600 frames from HD cameras 16 and 30 and for 4 scenarios: Haggling, Mafia, Ultimatum, Pizza. We train the model on the other 28 HD cameras of this dataset. Human 3.6M [23] is a dataset containing 3.6 million single-person RGB images with 3D human poses annotated by MoCap systems. We used the standard protocol for the evaluation. We used the S1, S5, S6, S7 and S8 subjects for training and the subjects S9 and S11 for testing.

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. The whole training procedure took about five days. We used the Adam[24] optimiser with an initial learning rate of 10^{-4} decreased to 10^{-5} at the 70000^{th} iteration, to 10^{-6} at the 200000^{th} and to 10^{-7} at the 220000^{th}.

Multi-person 3D human pose estimation results: Table 1 provides results of our method on the CMU-Panoptic dataset for 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 final model (last row), with four hourglass modules and the readout procedure described in Section 3.2 improves the results over the recent state of the art methods. Note that unlike [20] 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.

Single-person 3D human pose estimation results : Table 2 provides results of our method on the Human 3.6M dataset. While designed for multi-person 3D human pose estimation, our model produces reliable results in a single person setting with an MPJPE of 66.4 mm on the Human 3.6M dataset, better than most compared approaches. In particular, it has a lower error than [22] that also uses ORPM but differs in the architecture used and in the joint grouping method.

Method	Haggling	Mafia	Ultimatum	Pizza	Mean
[14]	217.9	187.3	193.6	221.3	203.4
[20]	140.0	165.9	150.7	156.0	153.4
[21]	72.4	78.8	66.8	94.3	72.1
Ours, 1-HG	92.3	86.1	82.7	103.8	91.8
Ours, 2-HG	77.1	74.8	68.0	89.8	78.3
Ours, 3-HG	72.4	72.4	60.12	85.2	73.8
Ours, NR	101.5	124.2	105.7	130.3	118.8
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. (i-HG stands for i stacked hourglasses).

	Direction	Discussion	Eating	Greet	Phone	Photo	Pose	Purchase
[11]	**50.1**	**54.3**	57.0	**57.1**	66.6	73.3	53.4	55.7
[12]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3
[16]	54.8	60.7	58.2	71.4	**62.0**	**65.5**	53.8	55.6
[10]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
[13]	62.6	78.1	63.4	72.5	88.3	93.8	63.1	74.8
[3]	52.5	63.8	**55.4**	62.3	71.8	79.8	**52.6**	72.2
[19]	76.2	80.2	75.8	83.3	92.2	79.9	105.7	71.7
[22]	58.2	67.3	61.2	65.7	75.82	84.5	62.2	64.6
Ours	**50.1**	66.4	56.4	65.0	69.4	81.5	55.6	**52.1**
	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	AVG
[11]	**72.8**	**88.6**	**60.3**	**57.7**	62.7	**47.5**	**50.6**	**60.4**
[12]	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
[16]	75.2	111.6	64.2	66.1	**51.4**	63.2	55.3	64.9
[10]	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
[13]	106.6	138.7	93.8	73.9	82.0	55.8	59.6	80.5
[3]	86.2	120.6	66.0	64.0	76.8	48.9	53.7	68.6
[19]	105.9	127.1	88.0	83.7	86.6	64.9	84.0	87.7
[22]	82.0	93.0	68.8	65.1	72.0	57.6	63.6	69.9
Ours	83.8	115.4	62.7	64.4	78.1	48.0	53.1	66.4

Table 2: Mean per joint position error (MPJPE) in mm on the Human3.6M dataset.

5. CONCLUSION

We have presented a single shot trainable model for multi-person 3D human pose estimation in various camera viewpoint conditions, strong occlusions and various social activities. 2D and 3D human joints are predicted using heatmaps and ORPM which have proven their ability to manage 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 provided experiments 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.

6. REFERENCES

[1] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.

[2] Alejandro Newell, Zhiao Huang, and Jia Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *NIPS*, 2017.

[3] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3D Vision*, 2017.

[4] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al., "Panoptic studio: A massively multiview system for social interaction capture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 190–204, 2019.

[5] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[6] Alexander Toshev and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *CVPR*, 2014.

[7] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy, "Towards accurate multi-person pose estimation in the wild," in *CVPR*, 2017.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *ICCV*, 2017.

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," 2017.

[10] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little, "A simple yet effective baseline for 3d human pose estimation," in *ICCV*, 2017.

[11] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation," 2018.

[12] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *CVPR*, 2017.

[13] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, 2017.

[14] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu, "Deep multitask architecture for integrated 2d and 3d human sensing," in *CVPR*, 2017.

[15] Grégory Rogez and Cordelia Schmid, "Mocap-guided data augmentation for 3d pose estimation in the wild," in *NIPS*, 2016.

[16] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *ICCV*, 2017.

[17] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis, "Ordinal depth supervision for 3d human pose estimation," 2018.

[18] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang, "Adversarial learning of structure-aware fully convolutional networks for landmark localization," 2017.

[19] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid, "Lcr-net: Localization-classification-regression for human pose," in *CVPR*, 2017.

[20] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu, "Monocular 3d pose and shape estimation of multiple people in natural scenes–the importance of multiple scene constraints," in *CVPR*, 2018.

[21] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu, "Deep network for the integrated 3d sensing of multiple people in natural images," in *NIPS*, 2018.

[22] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt, "Single-shot multi-person 3d body pose estimation from monocular rgb input," 2017.

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *PAMI*, vol. 36, no. 7, 2014.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.