JMLR: Workshop and Conference Proceedings 101:48–63, 2019 ACML 2019

An Articulated Structure-aware Network for 3D Human Pose Estimation

Zhenhua Tang TANGEZREAL@GMAIL.COM
College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China.

Xiaoyan Zhang* XYZHANG15@SZU.EDU.CN
College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China.

Junhui Hou JH.HOU@CITYU.EDU.HK
Department of Computer Science, City University of Hong Kong, Hong Kong, China.

Abstract

In this paper, we propose a new end-to-end articulated structure-aware network to regress 3D joint coordinates from the given 2D joint detections. The proposed method is capable of dealing with hard joints well that usually fail existing methods. Specifically, our framework cascades a refinement network with a basic network for two types of joints, and employs a attention module to simulate a camera projection model. In addition, we propose to use a random enhancement module to intensify the constraints between joints. Experimental results on the Human3.6M and HumanEva databases demonstrate the effectiveness and flexibility of the proposed network, and errors of hard joints and bone lengths are significantly reduced, compared with state-of-the-art approaches.

Keywords: 3D human pose estimation, Articulated structure-aware network, Attention module, Random enhancement module.

1. Introduction

3D human pose estimation (3D-HPE) is a fundamental problem of computer vision, which has a broad spectrum of applications in entertainment, action recognition, and human computer interaction (Tu et al. (2018); Luvizon et al. (2018); Sminchisescu (2008)). With the great success of 2D human pose estimation (2D-HPE) (Newell et al. (2016); Wei et al. (2016); Cao et al. (2017); de Bem et al. (2018)), a common strategy for 3D-HPE task is lifting 2D joint detections into 3D space. However, most lifting based methods mainly focus on reducing the overall joint errors and ignore the special topological structure of human, which may induce in large errors for some joints with larger motion space. Here, we discuss two categories of the most related works: camera based model and deep learning based model.

Camera Based Model: This group of methods build a geometric model of camera, and optimize the camera parameters by matching 2D poses with the corresponding 3D poses (Wang et al. (2014); Zhou et al. (2015, 2017, 2018)). However, camera based methods heavily depend on an ideal geometric model (Wang and Wu (2011)).

* Corresponding author

© 2019 Z. Tang, X. Zhang & J. Hou.

Figure 1: **Motion spaces of joints.** The left image shows the motion locations of a hard joint (left wrist) in a video, while the right image shows the motion locations of a non-hard joint (left shoulder) in the same video. The length of the video is 27 seconds and it has in total of 1382 frames. Both the two screens in the images are the first frame of the video. std_x is the standard deviation of joint coordinates in x-axis, while std_y for the one in y-axis.

Deep Learning Based Model: Benefitting from the strong fitting power of deep neural network, deep learning based methods achieve superior performance for 3D-HPE. Considering the fact that distances of paired joints are usually distributed in a more compact space, Moreno-Noguer (2017) regressed a 3D distance matrix of joints by using a fully convolutional network. Martinez et al. (2017) designed a simple fully connected network and further encouraged the exploration of the network architectures for lifting 2D to 3D. Furthermore, Pavlakos et al. (2018) added the ordinal depth information as input for the ambiguity of 3D-HPE, and Ronchi et al. (2018) introduced a ranking loss to relieve the limited availability of annotations.

However, the methods mentioned above mainly focused on reducing the overall joint errors and not fully considered the topological structure of human body. Those existing approaches generally estimate the 3D human pose by holistic prediction, which suffer from poor performance for locating challenging joints. We divide the joints in the human body into hard joints (i.e., elbows, wrists, knees and feet) and non-hard joints (joints excluding hard joints). The intuition behind such a division of body joints is based on the estimation difficulty (i.e., the estimation error between the predicted coordinates and the real coordinates).

There are two main reasons why the hard joints show much larger errors than the non-hard ones in 3D-HPE: (1) the positions of hard joints can vary in a much lager space than the non-hard joints, leading that the standard deviations of hard joint coordinates are

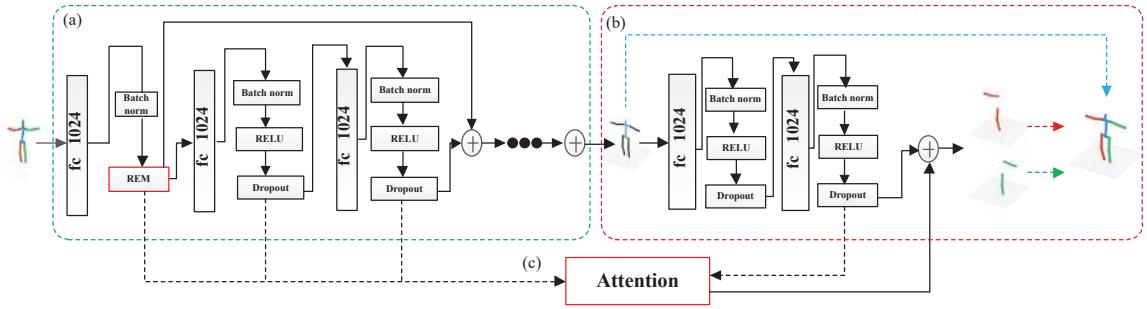

Figure 2: **Overall architecture of the proposed model.** The proposed network contains three parts: (a) a basic network with the proposed REM, (b) a refinement network and (c) the proposed AM.

much lager than non-hard joint coordinates, as depicted in Fig.1; and (2) due to the mutual inferences of joints, the errors of joints may be transferred and accumulated at hard joints.

Motivated by the above observations, in this paper, we propose a human structure-aware fully connected network for the 3D-HPE task. We cascade a basic network with a refinement network, where the former is used to regress a coarse 3D human pose and the latter is used to relocate the hard joints. Additionally, we propose an attention module (AM) and a random enhancement module (REM) to improve the ability of network fitting. The AM reliefs the propagations of errors, and the REM promises a strong relations between joints. We verified the effectiveness and flexibility of our network by a series of experiments on two publicly available Human3.6M and HumanEva databases.

In general, our contributions can be summarized as follows:

- We propose a human structure-aware network to utilize the body structure information for 3D-HPE, which outperforms most existing methods without any additional input information on two commonly-used databases.
- We formulate an AM to filter the desired features for hard joints refinement, and it enhances the learning of the fully connected network.
- We design a REM to augment the ability of the network the specific relationships between joints.

2. Proposed Method

In this section, we present different modules of our proposed network for 3D-HPE, as shown in Fig.2. The network consists of three parts, a basic network with the REM (in (a)), a refinement network (in (b)) and an AM (in (c)). First, we lift a 2D human pose to a coarse 3D pose, relying on a simple basic network. Meanwhile, a REM is presented to explore the constraints between joints for 3D human pose. Then, we formulate an AM to weight features and filter the desired ones. Finally, we adjust the hard joint locations by proposing a refinement network.

2.1. Basic network

Our initial goal of the basic network is to regress a coarse 3D human pose. We choose the network proposed by Martinez et al. (2017) for its efficiency in 3D-HPE. This network consists of multiple stacked residual modules, where each module contains two linear fully connected layers with the dimension of 1024. We use three black round solid dots to present more residual modules in Fig.2 (a). Given n 2D joints $\mathbf{x} \in \mathbb{R}^{2n}$ and their corresponding 3D joints $\mathbf{y} \in \mathbb{R}^{3n}$, the basic network aims to learn a function $f_b : \mathbf{x}^{2n} \rightarrow \mathbf{y}^{3n}$ that regresses 3D joint locations with the 2D inputs. The loss function is

$$L_1 = \sum_{i=1}^{n} \mathcal{L}(f_b(\mathbf{x}_i) - \mathbf{y}_i), \tag{1}$$

where $\mathcal{L}$ is $\mathcal{L}_2$-Norm.

We follow the same settings of the original network and refer the interested reader to literature Martinez et al. (2017), where those contributions have been demonstrated.

2.2. Refinement network

The basic network doesn't work well for hard joints, since all joints share the same neurons in the network. Moreover, the motion spaces of hard joints are much lager than those of non-hard joints, hence more parameters are required to lift the 2D hard joints. One solution is to deepen the network, however, it may result in overfitting for the non-hard joints.

Considering the structural characteristics of human body, we cascade a refinement network with the basic network to refine the hard joints. The refinement network employs multiple fully connected layers with the dimension of 1024. We first obtain a set of 3D locations from the basic network, including coarse locations of hard joints and precise locations of non-hard joints. Then, we feature and propagate all the 3D joint locations to our refinement network. Finally, we explicitly learn accurate locations for hard joints given preliminary coarse joint coordinates.

By supervising the final layer of the basic network, the results of joints from the basic network stay the same level with those of the network in literature Martinez et al. (2017). In order to avoid overfitting caused by the refinement network and ensure an end-to-end learning pattern, the refinement network only back propagates the loss of hard joints. Let $\mathbf{q}$ denote the indication, the refinement network learns a function f_r to refine the hard joints with a loss function as follows:

$$L_2 = \sum_{i=1}^{n} \mathbf{q}_i \times \mathcal{L}(f_r(f_b(\mathbf{x}_i)) - \mathbf{y}_i), \tag{2}$$

where $\mathcal{L}$ is $\mathcal{L}_2$-Norm and

$$\mathbf{q_i} = \begin{cases} 1, & if\ the\ i_{th}\ joint\ is\ a\ hard\ joint, \\ 0, & otherwise. \end{cases}$$

Finally, we formulate the ultimate objective function to minimize the regression error as:

$$L = \min(L_1 + L_2), \tag{3}$$

where $\mathcal{L}$ is $\mathcal{L}_2$-Norm.

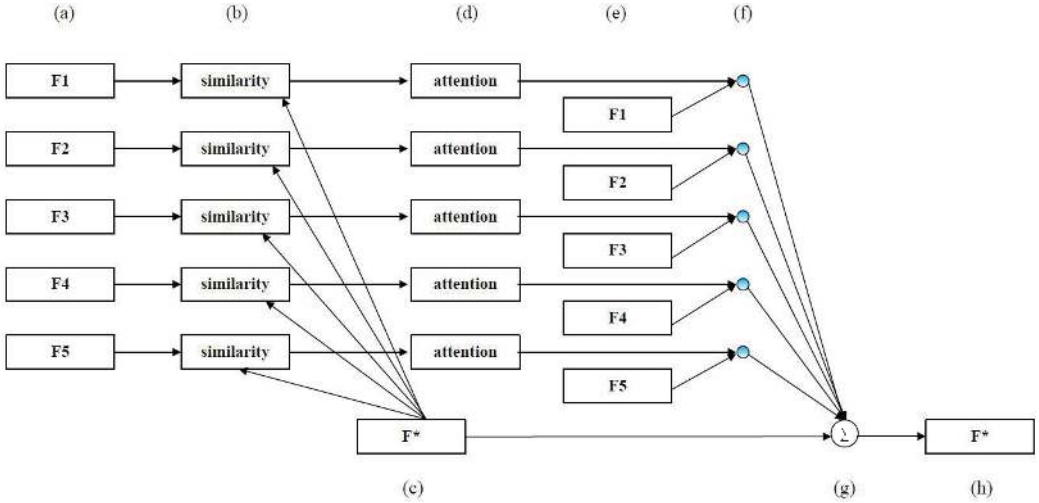

Figure 3: **The architecture of the AM.** (a) and (e) are the 2D features. (c) is the original 3D feature. (b), (d) and (g) indicate respectively the calculation of Eq.(6), Eq.(7) and Eq.(8). (f) represents the dot product operation. (h) is the final output feature.

2.3. Attention module

In geometry, 3D pose reconstruction heavily requires the corresponding 2D information (Wang and Wu (2011)). The camera pose estimation aims to estimate the extrinsic parameters i.e., the rotation matrix R and the translation vector $\mathbf{t}$, and possibly all or a subset of the intrinsic parameters K. K is expressed as

$$K = \begin{bmatrix} \alpha f & s & u \\ 0 & f & v \\ 0 & 0 & 1 \end{bmatrix}, \tag{4}$$

where f denotes the focal length, (u, v) denotes the principal point, α is an aspect ratio, and s is the skew. Given 2D and 3D point correspondences, X and Y, the relationships for 2D and 3D joints admit the following projection equation:

$$X = K[R, t]Y. \tag{5}$$

For our refinement network, only utilizing the coarse 3D pose as input may result in a serious weakening of geometrical relationship between 2D and 3D. This geometric relationship can be reflected by matching 2D and 3D. Therefore, we propose to fuse features from the basic network to the refinement network to enhance the geometry information for hard joints, through a fully connection operation. This approach is similar to implicitly building a camera projection model, which ensures the 2D information to be transfered to the refinement network. However, directly connecting the 2D features and the 3D features may result in information interference. To address this problem, we introduce an effective method for feature fusion in the following paragraphs. Since the basic network inputs are only 2D joint

locations, we named the features in the basic network as 2D features. Similarly we named the features in the refinement network as 3D features.

One of the most key aspects of the human visual system is the presence of attention, which allows the system to focus on salient regions. The attention mechanism can thus be beneficial to the extraction of the informative features obtained from the basic network. Though the utilization of attention has proved to be an effective solution toward image understanding (Zhao et al. (2017); Wang et al. (2017)) or natural language processing (Yin et al. (2016); Yang et al. (2016)) in previous works, there are no rigorous studies yet on how to integrate the attention into 3D-HPE. Therefore, a principle way is needed to learn a model that can attend to the desired 2D features when refines the hard joints.

In this paper, we develop an AM to model the attention on 2D features, as shown in Fig.3. Technically, given 2D features $\mathbf{F}_i$ and a 3D feature $\mathbf{F}^*$ from the last layer of the refinement network, we feed them into the proposed AM to generate the attention distribution. All the selected features are from the layers after dropout and have same dimension of 1024. We don't transform the inputs features when forward-propagate them to the AM, so we represent the inputs procedures as dotted lines as shown in Fig.2 (c). The process is formatted as follow:

$$\mathbf{sim}_i = \mathbf{w}_i \times \mathcal{L}(\mathbf{F}_i - \mathbf{F}^*), \tag{6}$$

$$\mathbf{att}_i = 1 - sigmoid(\mathbf{sim}_i), \tag{7}$$

$$\mathbf{F}^* = \mathbf{F}^* + \sum_{i=1}^{m} \mathbf{att}_i \times \mathbf{F}_i, \tag{8}$$

where $\mathbf{w}_i$ is a parameter vector and $\mathcal{L}$ is $\mathcal{L}_2$-Norm. This AM is a differentiable expression. The Eq. (6) is employed to weight the similarities between features, and the similarities are normalized to $[0, 1]$ as attentions in Eq. (7). The Eq. (8) assigns the 2D information to 3D features according to different attention weights.

In general, our AM filters out the distracting information and enforces the geometry relations of 2D and 3D. It improves the fitting ability of the whole network. If we directly connect the final 2D feature and the final 3D feature, without the AM, the refinement network unable to obtain optimal information and optimal 3D human pose (the AM* in Table 2). Importantly, our AM not only applies to the proposed network, it can be extended to other fully connected network for desired information by modifying the similarities conditions in the Eq. (6).

2.4. Random Enhancement Module

In the basic network, Dropout (Srivastava et al. (2014)) is employed after RELU (Glorot et al. (2011)) to prevente co-adaptation of feature detectors, for significantly reducing overfitting (Hinton et al. (2012)). According to literature Goodfellow et al. (2016), the main power of dropout comes from the fact that the making noise is applied to the hidden units. This can be thought of as a form of highly intelligent and adaptive destruction of input information. For example, in 3D-HPE, if the model learns a neuron h_i that locating wrist by elbow, then dropping h_i is equal to erase the constraint information of the forearm. The

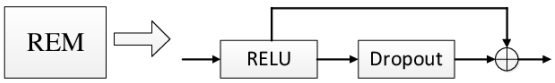

Figure 4: **The architecture of the REM.** The REM connects a RELU layer with a Dropout layer.

model must learn another neuron h_j, that either redundantly encodes the presence of elbow, or locates the wrist by other features. Hence, when we gradually adjust the Dropout from 0.5 to 1.0 with a step of 0.1, the structural relationships of body are preserved, and the error of bone lengths decreases as the number of retained neurons increases. However, the joint errors are increased due to overfitting (More detailed explanation is referred to the Section 5.2 of Supplementary.).

To address this problem, we propose to introduce a REM to replenish the joint information for bone length computation without parameter increasing, which may reduce the impact of joint constraint vanish caused by dropout. The architecture of the REM simply consists of RELU and Dropout, these two operations are wrapped in a residual connection, as shown in Fig.4. Such that, the randomness of Dropout is preserved for preventing co-adaptation of feature detectors. And we keep the integrity of features for strong constraint relationship. Intuitively, the REM achieves the effect of random enhancement of information. The REM is used only at the first layer of the basic network, where there is the original source of information.

3. Experiments

This section concerns the empirical evaluations of the proposed approach. First, we present some implementation details of the proposed network. Then, quantitative and qualitative results are illustrated on the selected datasets.

3.1. Implementation Details

Dataset: We evaluate our proposed approach on two publicly available datasets: Human3.6M (Ionescu et al. (2014)) and HumanEva (Sigal et al. (2010)).

Human3.6M is a large-scale dataset captured in an indoor environment that contains 11 subjects performing 15 typical actions like eating, walking, making a phone call and engaging in a discussion. We follow the standard protocol of Human3.6M, which is using subjects 1, 5, 6, 7, and 8 for training, and subjects 9 and 11 for evaluation. We report the average errors in millimetres between the ground truth and our predicted 3D locations across all joints.

HumanEva is a small database that has been largely used in many benchmark works over the last decade. The dataset contains 4 subjects performing a 6 common actions (e.g. walking, jogging, gesturing, etc.).

2D detection: Unless stated otherwise, the Stacked Hourglass Network (Newell et al. (2016)) is adopted to obtain 2D joint locations, which achieves sufficient accuracy for our 3D-HPE task, and pre-trains on the MPII dataset.

Table 1: **Results on the Human3.6M.** Comparison of overall average errors (mm) for different methods using the Human3.6M dataset. '-' means that the result of corresponding work is not reported.

Methods	Direct	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Ionescu et al. (2014)	132.71	183.55	133.37	164.39	162.12	205.94	150.61	171.31
Li et al. (2015)	-	138.88	96.94	124.74	-	168.08	-	-
Tekin et al. (2016)	-	129.06	91.43	121.68	-	-	-	-
Zhou et al. (2016)	87.36	109.31	87.05	103.16	116.18	143.32	106.88	99.78
Moreno-Noguer (2017)	67.44	63.76	87.15	73.91	**71.48**	69.88	65.08	**71.69**
Sun et al. (2017)	90.20	95.50	82.30	85.00	87.10	87.90	93.40	100.30
Martinez et al. (2017)	52.99	60.55	62.08	62.72	86.10	57.96	58.89	81.70
Zhou et al. (2018)	68.70	74.80	67.80	76.40	76.30	84.00	70.20	88.00
Ours	**51.77**	**59.38**	**59.55**	**60.33**	84.15	**56.22**	**57.36**	79.73
Methods	SitDown	Smoke	Photo	Wait	Walk	WalkD	WalkT	Ave.
Ionescu et al. (2014)	151.57	243.03	162.14	170.69	177.13	96.6	127.88	162.1
Li et al. (2015)	-	-	-	-	132.17	69.97	-	-
Tekin et al. (2016)	-	-	162.17	-	65.75	130.53	-	-
Zhou et al. (2016)	124.52	199.23	107.42	118.09	114.23	79.39	97.7	112.91
Moreno-Noguer (2017)	98.63	81.33	93.25	74.62	76.51	77.72	74.63	76.47
Sun et al. (2017)	135.40	91.40	94.50	87.30	78.00	90.40	86.50	92.40
Martinez et al. (2017)	99.59	69.17	83.49	64.04	50.72	67.49	54.63	67.48
Zhou et al. (2018)	113.80	78.00	98.40	90.10	62.60	75.10	73.60	79.90
Ours	**97.68**	**67.07**	**80.44**	**62.53**	**48.76**	**65.25**	**52.48**	**65.51**

Training details: The basic network is realized by 2 residual modules total 5 fully connected layer in dimension 1024, and the refinement network consists of 2 layers. We initialize the parameters using Kaiming initialization and use Adam to optimize the network. In the process of training, a mini-batch size is set to 64 for 300 epoches on Human3.6M (200 epoches on HumanEva). The learning rate is initialized as 1×10^{-3} with exponential decay, and the Dropout is set to 0.5. The proposed network is trained on a server with dual physical cores (Intel Xeon CPU E5-2690 v4 2.60GHz), one piece of GPU(Tesla P100-PCIE-16GB) and 256GB main memory.

3.2. Quantitative Results on Human3.6M

Comparison with related work. First, we examine the quantitative results of our human structure-aware network. Following previous works, the detailed results in terms of the Root Mean Square Error (RMSE) are presented in Table 1. Our results are also compared with those of related methods in Table 1. Unsurprisingly, making use of the structural information of human body helps for 3D-HPE. Without any additional input information, our human structure-aware network network achieves state-of-the-art results across most of actions, that the absolute error reduces about $2mm$ on average comparing with our baseline work Martinez et al. (2017). Since our work aims at reducing the errors of hard joints, most of the improvement for our approach comes from adjusting the hard joints, so improvements in all actions are similar , about $1.3mm$ to $2.5mm$, comparing with our baseline work Martinez et al. (2017). Specially, work Moreno-Noguer (2017) explicitly

constraints distances between joints to normalize all actions to a compact space, and gets smaller errors on poses with high complexity, i.e. Phone and Sitting. This illustrates the importance to consider about the human structure for complex pose estimations. However, the errors of other actions by work Moreno-Noguer (2017) are larger than those of our method.

Table 2: **Ablation studies of our network.** 'ref' means that there are no 2D features propagating to the refinement network. 'AM*' means that we connect the final layer of the basic network with the refinement network (without the AM). 'AM' indicates the Attention Module , and 'REM' is the Random Enhancement Module.

module	average error
basic	67.48
basic+ref	68.66
basic+ref+AM*	66.78
basic+ref+AM	66.15
basic+ref+AM+REM	65.51

Ablation studies of our network. Further, we present the ablation studies in Table 2, to demonstrate the significance of each module in the network. If we directly cascade the refinement network with the basic network, the hard joints are equivalent to meaningless search in a larger and challenge space without geometrical relationship between 2D and 3D. And our network is easier to get caught up in overfitting than the basic network. Hence, the error of term 'basic+ref' ($68.66mm$) is worse than the term 'basic' ($67.48mm$). After inducing 2D location cues by connecting the final layer of the basic network with the refinement network, the refinement network relocates hard joints to geometrical relation constrained positions. Therefore the error of term 'basic+ref+AM*' ($66.78mm$) reduces the error by about $0.7mm$ comparing with the basic network. However, crudely introduce the 2D information by a fully connection may result in interfering information, that is should be avoided. Benefiting from the AM, the refinement network learns and selects more effective features such that the hard joints relief from the propagation of error, and the error is further reduced to $66.15mm$. Finally, the REM enhances the collaboration ability of neurons to intensify the constraints of joints, so the network achieves the best results, with $65.51mm$.

Errors of joins. To demonstrate the significant efficiency of our network, the errors of hard joints are separately analyzed and compared in Fig.5 (a). Benefiting from multiple modules in the proposed network, the errors of all hard joints are significantly reduced by $2.27mm$ to $7.45mm$ comparing with the errors of basic network. For some challenging joints with largest motion spaces and strongest interferences, such as the wrists, the errors of them are reduced more significantly (more than $6mm$). This demonstrates the effectiveness of our human structure-aware network on reducing errors of hard joints. However, the 3D hard joints can be further optimized as they still have large errors. Note that, the error is mainly due to the corresponding 2D hard joint detections. Moreover, we also report the

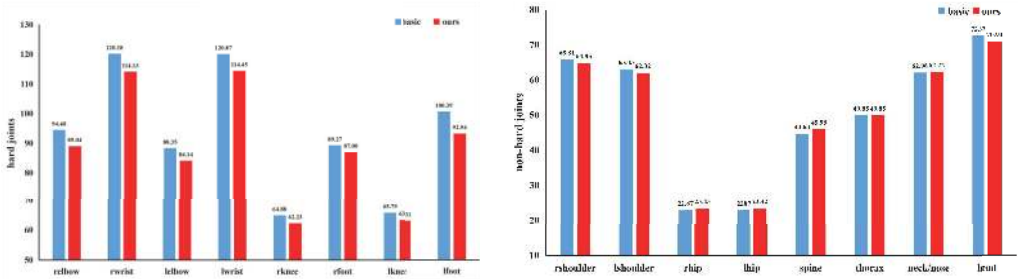

Figure 5: **Detailed results of joints.** (a) and (b) respectively indicate the errors (in mm) of hard joints and non-hard joints. The blue bars represent the joint errors of the basic network, and the red bars correspond to results of the proposed network. 'r' represents the joints on the right of human body, and 'l' represents the joints on the left of human body.

non-hard joint errors in Fig.5 (b), to show that our refinement network has little even no effect on the non-hard joints.

Table 3: **Errors of limb lengths.** RMSE (mm) of the bone lengths. 'lu-arm' denotes the upper left arm and 'll-arm' for the lower left arm. Similarly, 'ru' denotes the upper right ones and 'rl' for the lower right ones.

	basic	ours	reduction rate (%)
lu_arm	16.90	14.19	11.60
ll_arm	28.29	20.64	27.00
lu_leg	17.21	16.06	6.68
ll_leg	27.15	23.40	13.81
ru_arm	16.45	13.66	16.96
rl_arm	29.99	21.27	29.08
ru_leg	16.62	16.38	1.44
rl_leg	26.98	22.70	15.86
avg	22.45	18.53	17.42

Analysis of bone length errors. Furthermore, the advantage of our method on reducing the bone length errors is also analyzed, and the results are reported in Table 3. Comparing with the basic network, our approach estimates a more reasonable 3D human pose whose limbs length errors reduce significantly by 17%. This effect on reducing bone length errors is a benefit from accurate coordinates of hard joints. On the other hand, we retain more information about the structure of the human body in our network by using the AM and REM. Therefore, our method gets more precise limb length.

Expending experiment about depth order. Finally, an expending experiment result is shown in Table 4 to prove the potential of our network. Following the literature

Table 4: **Expending experiment.** We use the gaussian noise to interfere depth values, '-' means that the value of corresponding work is not reported.

module	depth noise	average error (mm)
Pavlakos et al. (2018)	-	41.80
basic	$\mathcal{N}(0,0.85)$	41.51
ours	$\mathcal{N}(0,0.85)$	38.94

Pavlakos et al. (2018), we use accurate 2D joint locations and a noisy version of the depth order of the joints as input, such that the majority of their ordinal relations are preserved. After inducing ordinal relations of joints, all the three models in Table 4 release from serious ambiguity of 3D-HPE. At the gaussian noise with average value of 0 and variance of 0.85, the basic network achieves the same performance of average joint error with the literature Pavlakos et al. (2018). However, both the literature Pavlakos et al. (2018) and the basic network can not deal with the hard joints. Hence, we refine the hard joints by the proposed network. After fully considering the structural characteristics of the human body, our network further reduces the error to 38.94mm.

3.3. Quantitative Results on HumanEva

Table 5: **Results on the HumanEva.** Comparison of overall average errors (mm) for different methods using the HumanEva dataset. 'gt' means that we use ground truth of 2D pose as input for the network.

	S1		S3	
	walking	jogging	walking	jogging
Radwan et al. (2013)	75.1	79.2	93.8	99.4
Wang et al. (2014)	71.9	62.6	85.3	54.4
Simo-Serra et al. (2013)	65.1	74.2	73.5	32.2
Bo and Sminchisescu (2010)	46.4	64.5	64.9	38.2
Kostrikov and Gall (2014)	44.0	57.2	41.7	40.3
Yasin et al. (2016)	35.8	46.6	41.6	38.9
Moreno-Noguer (2017)	19.7	39.7	24.9	21.0
Pavlakos et al. (2017)	22.1	29.8	29.0	26.0
basic_gt Martinez et al. (2017)	17.58	26.18	30.40	20.93
Ours_gt	**16.39**	**22.78**	**26.84**	**17.96**

Comparison with related work. We report the RMSE (mm) on the HumanEva dataset in Table 5. We use the 2D pose ground truth as input of the baseline Martinez et al. (2017) and our model. Since HumanEva is a small dataset without complex 3d human pose, most related works get rather small errors than those on Human3.6M. By refining the locations of hard joints, our network significantly reduces the errors of relatively difficult

actions (i.e., jogging in S1 and walking in S3). In this case, our network still achieves the best performance in all subjects and actions. Our network shows strong stability and generalization, even on different datasets.

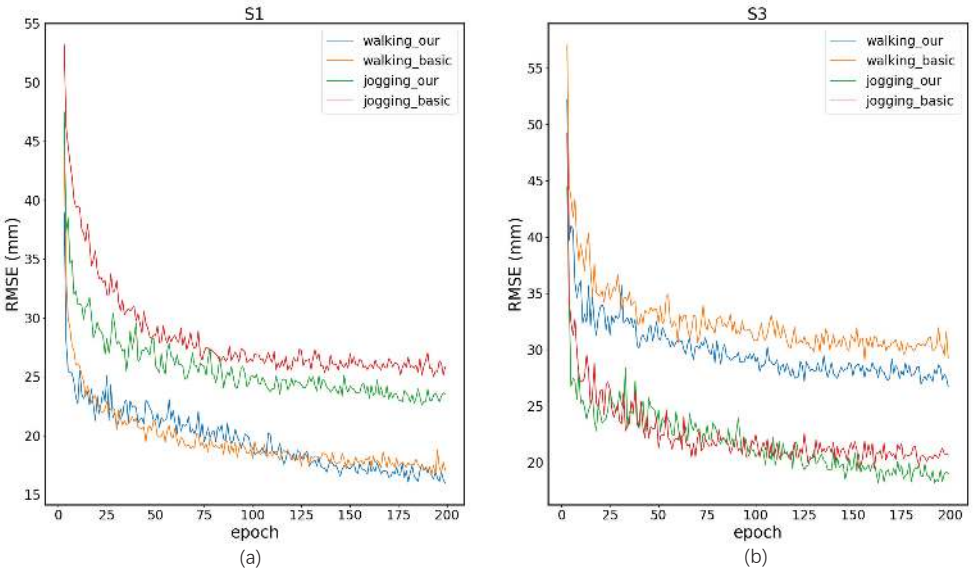

Figure 6: **Convergence analysis of networks.** The horizontal axis represents the numbers of networks training iterations, and the vertical axis represents the RMSE of the models in the test set. (a) and (b) respectively correspond to result on two subjects S1 and S3.

Convergence analysis of networks. We further analyze the convergence effect of our model. As shown in the Fig.6, the proposed method outperform the baseline in both convergence speed and error. For action walking on S1 and action jogging on S3, due to the low complexity of the pose, the RMSEs of both networks cross fluctuation in the first half of the training, our model has no significant effect. However in the second half of training, when the network of baseline converges to relatively large errors (yellow line in (a) and red line in(b)), our deeper network continues to be optimized and gets small RMSEs (blue line in (a) and green line in (b)). For action jogging on S1 and action walking on S3, our method is remarkably effective in adjusting difficult pose which suffer from poor performance for hard joints, and the RMSEs of our network (green line in (a) and blue line in (b)) are always less than the baseline's (red line in (a) and yellow line in (b)).

3.4. Qualitative Results

In Fig.7, we have collected a sample of 3D pose outputs in Human3.6M for our approach. As shown in the first lines, our network regresses a precise 3D pose for the upright human pose. For the complex pose like sitting on a chair as shown in the second line, even if there is a small error at the hard joints (i.e., left wrists), our arms still extend to a relative correct direction. In addition, results in the last column reveal some of the limitations of our

approaches, where the wrists and knees get into big trouble because of self-occlusions. This case prompts us to think more about the problem further, that how to utilize the structure of human body and the depth information to deal with the self-occlusions of poses.

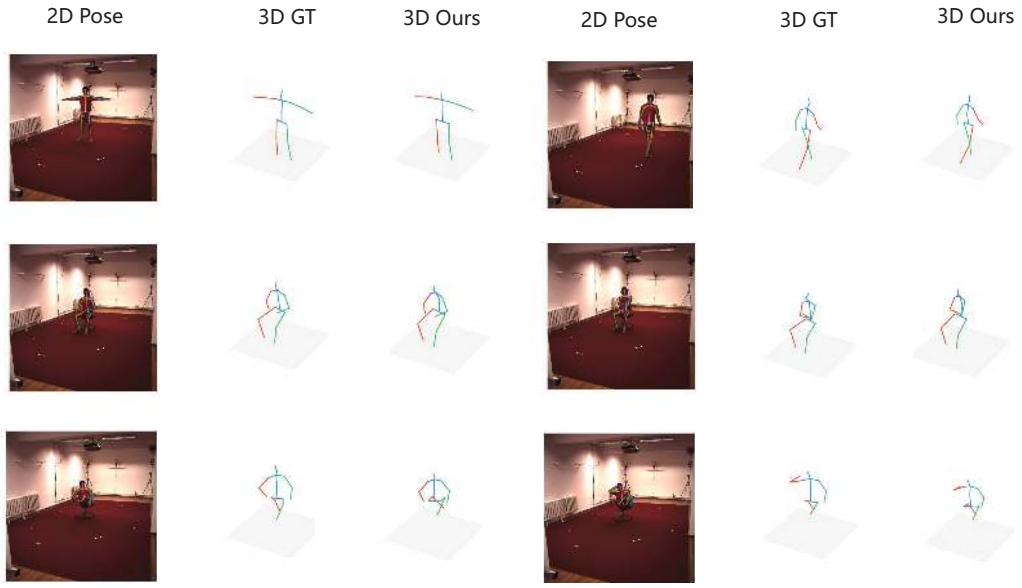

Figure 7: **Examples of selected poses using Human3.6M dataset.** The images from left to right in each triplet correspond to the given 2D pose (in a image), the grondtruth of 3D pose and the predicted 3D poses of our network, respectively. In the 3D poses, green is for the left side of the human body and red is for the right.

4. Conclusion

We have presented an effective solution for handling the hard joints in 3D-HPE, by constructing an articulated structure-aware network. Without any additional input information, we cascaded a refinement network with a basic network to adjust the hard joints. We formulate an AM to promise the propagation of the desired geometric information of 2D and 3D in the network. And we also improved the co-adaptation of neurons to enhance constraint relationships of joints, by proposing a REM. Extensive experimental results have demonstrated the superior performance of the proposed method to various classical and state-of-the-art ones, i.e., our proposed method can significantly reduce the errors of hard joins and bone length of limbs. We will further investigate the specific relations between joints to address the self-occlusion problem of 3D-HPE in future work.

Acknowledgements

The authors wish to acknowledge the financial support from:

(i) Chinese Natural Science Foundation under the Grant No.61602313;
(ii) Shenzhen Commission of Scientic Research & Innovations under the Grant No. JCYJ20170302153632883.

References

Liefeng Bo and Cristian Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1-2):28, 2010.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

Rodrigo de Bem, Anurag Arnab, Stuart Golodetz, Michael Sapienza, and Philip Torr. Deep fully-connected part-based models for human pose estimation. In *Asian Conference on Machine Learning*, pages 327–342, 2018.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian 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(7):1325–1339, 2014.

Ilya Kostrikov and Juergen Gall. Depth sweep regression forests for estimating 3d human pose from images. In *BMVC*, volume 1, page 5, 2014.

Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *IEEE International Conference on Computer Vision*, pages 2848–2856, 2015.

Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. *arXiv preprint arXiv:1802.09232*, 2018.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision*, volume 206, pages 2659–2668, 2017.

Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1561–1570, 2017.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017.

Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018.

Ibrahim Radwan, Abhinav Dhall, and Roland Goecke. Monocular image 3d human pose estimation under self-occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1888–1895, 2013.

Matteo Ruggero Ronchi, Oisin Mac Aodha, Robert Eng, and Pietro Perona. It's all relative: Monocular 3d human pose estimation from weakly supervised data. *arXiv preprint arXiv:1805.06880*, 2018.

Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

Edgar Simo-Serra, Ariadna Quattoni, Carme Torras, and Francesc Moreno-Noguer. A joint model for 2d and 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3634–3641, 2013.

Cristian Sminchisescu. 3d human motion analysis in monocular video: techniques and challenges. In *Human Motion*, pages 185–211. Springer, 2008.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017.

Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016.

Juanhui Tu, Mengyuan Liu, and Hong Liu. Skeleton-based human action recognition using spatial temporal 3d convolutional neural networks. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.

Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille, and Wen Gao. Robust estimation of 3d human poses from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2361–2368, 2014.

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

Guanghui Wang and QM Jonathan Wu. Simplified camera projection models. In *Guide to Three Dimensional Structure and Motion Factorization*, pages 29–41. 2011.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4956, 2016.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272, 2016.

Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6): 1245–1256, 2017.

Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, Kostas Daniilidis, et al. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 4447–4455, 2015.

Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4966–4975, 2016.

Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, and Kostas Daniilidis. Sparse representation for 3d shape estimation: A convex relaxation approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1648–1661, 2017.

Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.