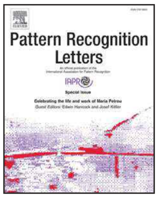
Pattern Recognition Letters 125 (2019) 404–410

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



3D human pose estimation via human structure-aware fully connected network

Xiaoyan Zhang[a,*], Zhenhua Tang[a], Junhui Hou[b], Yanbin Hao[b]

[a] College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China
[b] Department of Computer Science, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 11 August 2018
Available online 30 May 2019

MSC:
41A05
41A10
65D05
65D17

Keywords:
3D human pose estimation
Human structure
Fully connected network

ABSTRACT

Existing 3D human pose estimation (3D-HPE) methods focus on reducing the overall joint error, resulting in endpoints and bone lengths with large errors. To address this issue, we propose a human structure-aware network, which is capable of recovering 3D joint locations from given 2D joint detections. We cascade a refinement network with a basic network in a residual learning manner, meanwhile fuse the features from 2D and 3D coordinates by a residual connection. Specifically, our refinement network employs a dual-channel structure, in which the symmetrical endpoints are divided into two parts and refined separately. Such a structure is able to avoid the mutual interference of joints with large errors to promise reliable 3D features. Experimental results on the Human3.6M dataset demonstrate that our network reduces the errors of both endpoints and bone lengths compared with existing state-of-the-art approaches.

1. Introduction

3D human pose estimation (3D-HPE) has various applications such as virtual reality (VR), action recognition [3], and autonomous vehicles. Common strategy for obtaining 3D-HPE is inferring 3D joint locations from 2D joint detections, and this kind of methods can be roughly clarified into two categories: camera module based and deep learning based. Camera module based methods optimize the camera parameters in order to match the given 2D locations with their corresponding 3D representations [17]. However, these methods usually require an over-complete database [1,5] to cover various actions and also complex optimization methods [16,19–23]. Moreover, different camera models will have different effects on the results [18]. Deep learning based methods utilize the deep network models to recover 3D pose from given 2D locations, which achieve the state-of-the-art performance. Moreno–Nogue [11] regressed a 3D distance matrix via a Fully Convolutional Network(FCN) and then inferred the 3D poses by retrieving the joint locations which yield the same distance matrix. Martinez et al. [10] simply employed a multi-layer fully connected network to directly regress the 3D joint coordinates from 2D joint locations. It outperforms most existing methods and requires less time to train on Human3.6M.

All the above mentioned approaches deal with all joints equivalently, which concentrate on reducing the overall joint error, but ignore the articulated structure of the human body. The particular endpoints of such an articulated structure (i.e., elbows, wrists, knees and feet, namely **hard joints**) have a much larger motion space than the other joints (see Fig. 1), making the estimation more challenging. Meanwhile, the joint errors of limbs present an increasing trend due to the error propagation along this articulated structure. In addition, without involving in the relationships between body joints, these methods may result in unreasonable human poses, e.g., hand raising with a very long arm.

Motivated by the above observations, in this paper, we focus on designing a novel and deeper network model, which is to facilitate the 3D-HPE task. This proposed approach estimates 3D poses of body joints from a set of 2D joint locations via a human structure-aware fully connected neural network, which is a lying Y-shape network. The first part (front branch of the lying Y) of this network, a residual network (basic model), is for obtaining a set of 3D joint coordinate candidates. To exploit the structural information of human, we cascade the basic model with two individual fully connected networks (branches of Y) to refine the coordinates of hard joints. We named the two fully connected networks as refinement network. One is for refining the left hard joints (left-elbow, left-wrist, left-knee and left-foot), and the other is for the

[*] Corresponding author.
E-mail addresses: xyzhang15@szu.edu.cn (X. Zhang), jh.hou@cityu.edu.hk (J. Hou).

https://doi.org/10.1016/j.patrec.2019.05.020
0167-8655/© 2019 Elsevier B.V. All rights reserved.

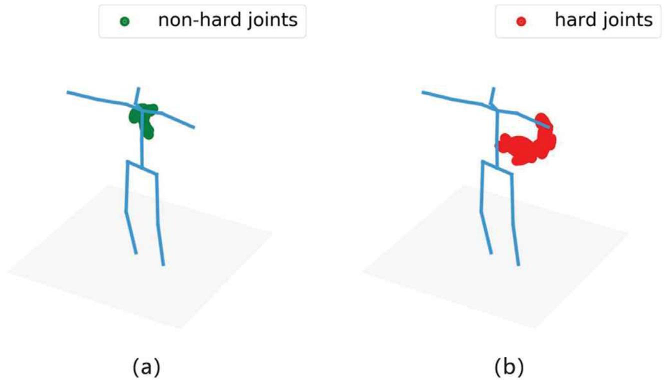


Fig. 1. Motion spaces of joints. The green dots in (a) and the red dots in (b) respectively represent the 3D coordinates of a non-hard joint (left shoulder) and a hard joint (left wrist), in 1552 frames of action sequences. This figure clearly shows that the hard joints have very larger motion space than non-hard joints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

right ones (right-elbow, right-wrist, right-knee and right-foot). This strategy avoids the inter-class influence between the left and right hard joints, and it enforces the inner-class relationship between the same group of hard joints. Furthermore, to reduce the loss of information during the translation from the basic network to the refinement network, we employ a residual connection between them. The effectiveness of the proposed approach are examined and compared against various state-of-the-art 3D-HPE approaches using the Human3.6M.

2. Proposed method

From Fig. 1, we can observe that the hard joints move in a much larger motion space than the non-hard joints. We need a deeper network to fit the large motion space of hard joints in the deep learning based methods. Simply deepening the network for 3D human pose estimation could reduce the length error of bones. However, it could not reduce joint errors. The joint error decreases first and then increases while increasing the network depth of a basic fully connected network, as depicted in our supplementary figure. Moreover, the joint error is mainly contributed by the hard joints. This observation inspires us that paying more attentions to hard joints in a deeper network benefits for reducing both joint errors and bone length errors.

The objective of this paper is to deepen the basic network to allow the network to capture more information from both 2D and 3D locations for reducing errors of hard joints and also bone length errors. We propose to first estimate the joint 3D coordinate candidates and then refine the locations of hard joints. To ensure end-to-end learning, we cascade a refinement network with a basic network to instead two separate networks. As the motion spaces of left endpoints have little relation with right endpoints, and the mutual inference between hard joints may aggravate the transfer error, therefore, we divide all the joints into the left and right sets for refinement. Finally, we cascade the basic model with two individual fully connected networks to refine the coordinates of hard joints.

Fig. 2 illustrates an overview of the proposed architecture. We first briefly review the structure of the basic network. Then a detailed discussion of our refinement network is presented.

2.1. Basic network

The basic network aims at directly regressing 3D joint locations from the 2D pose. This network consists of multiple stacked residual modules, each containing two linear fully connected layers of dimension 1024. Every two fully connected layers are wrapped in a residual connection. Particularly, every fully connected operation followed by Batch Normalization, RELU (Rectified Linear Units) [12] and Dropout [14]. With Kaiming initialization [2], given 2D joints $\mathbf{x} \in \mathbb{R}^{2n}$, and their corresponding 3D joints $\mathbf{y} \in \mathbb{R}^{3n}$, where n is the number of joints, the basic network aims to learn a function f_b: $\mathbf{x}^{2n} \rightarrow \mathbf{y}^{3n}$ that regresses 3D joints locations with a loss function as:

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

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

We choose this network for its efficiency for 3D-HPE [10], and it also takes less time to train in Human3.6M. However, all joints share the total neures in this network, and they have no explicit constraint among joints. This leads to large errors of hard joints and even a unreasonable pose with large bone length errors. For this issue, we explicitly refine the locations of hard joints and implicitly strengthen the relationship between joints by the proposed refinement network.

2.2. Refinement network

We first introduce a single-channel refinement network which aims at adjusting the hard joints for more precise coordinates, and then a dual-channel refinement network is introduced to further deal with the relationship between joints.

Single-channel: By supervising the final layer of the basic network, we propose to follow a coarse-to-fine learning pattern. We first obtain a set of 3D locations from the basic network, including coarse locations of hard joints and precise locations of other joints. Then we treat all the 3D joint locations as the feature being propagated to our refinement network. Finally, the refinement network explicitly learns accurate locations for hard joints given preliminary joint coordinates. The basic network and the refinement network are trained together in an end-to-end pattern using ground truth supervision. By doing so, the results of joints from the basic network stay the same level with those of the network in literature [10], even though the depth of the whole network is deeper. The architecture of the proposed single-channel refinement network is depicted in Fig. 3. To ensure an end-to-end learning pattern, we employ a binary vector as indication promising our refinement network only back propagating the loss of hard joints. Let $\mathbf{q}$ denote the indication, our refinement network learns a function f_r to refine the hard joints by 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

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

Dual-channel: Furthermore, we transform the above single-channel refinement network into a dual-channel module due to the special articulated structure of human body. In the above single-channel refinement network, 3D locations of all joints are considered as the input features of the refinement network. Differently, as shown in the Fig. 4, we divide human body joints into two partial overlapping categories, namely $A = \{left\_hand, left\_elbow, left\_knee, left\_foot, thorax, neck/nose, left\_shoulder, head, spine, left\_hip, pelvis\}$ and $B = \{right\_hand, right\_elbow, right\_knee, head, right\_foot, pelvis, neck/nose, right\_shoulder, thorax, spine, right\_hip\}$. In the learning phase, the network could implicitly take the constraint of bones into account. It explains why the bone errors decrease with the deeper network,

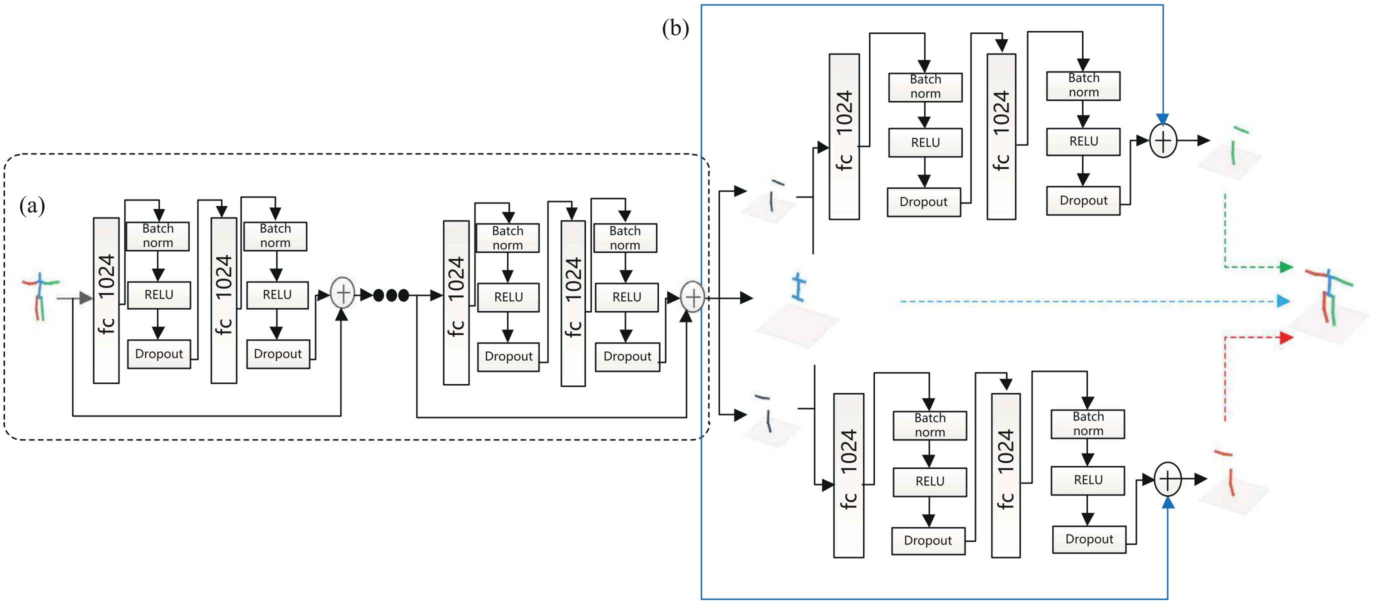


Fig. 2. Overall architecture of the proposed model. The proposed network contains two subnetworks: (a) a basic one and (b) a refinement one. The basic network consists of multiple stacked fully connected layers, and the refinement network possesses a dual-channel structure. The blue solid line indicates a residual connection. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

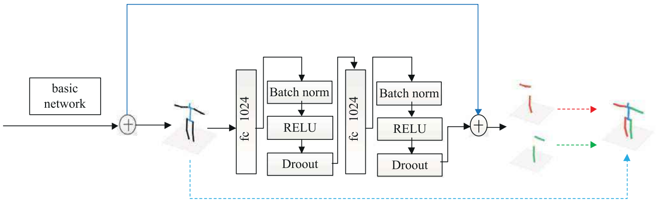


Fig. 3. Single-channel refinement network. The single-channel refinement network finetune the hard joint locations by using two fully connected layers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

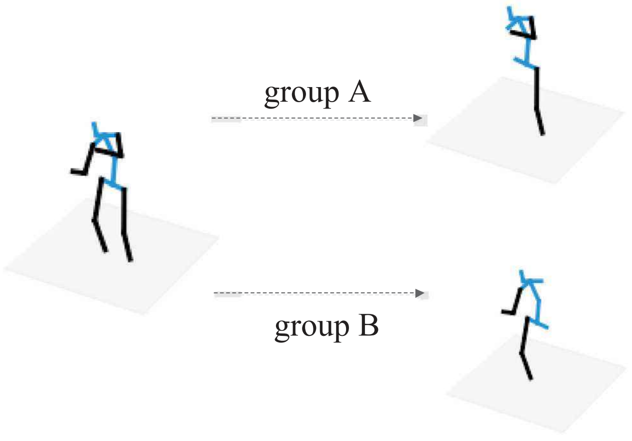


Fig. 4. Joint groups. The whole human body is divided into *A* and *B* group. Each group contains 4 hard joints (i.e. elbow, hand, knee and foot on black bones) and 2 non-hard joints (i.e. shoulder, hip on blue bones) on the left and right sides respectively, and shares all the rest joints (i.e. head, neck/nose, spine and thorax on blue bones). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

as is shown in Fig. 1 of supplementary. While refining the hard joints all together as in the single-channel refinement network, the large error hard joints of left offer wrong constraint information to the right hard joints.

In order to avoid the inter-class inference [7] between the left body and the right body hard joints and enforce the inner-class influence between the same group of joints, we employ two individual fully connected sub-networks to refine hard joints of left body and right body separately. Given obtained 3D joint coordinates by the basic network, we isolate *A* to be the input of the top channel, and finetune hard joints in *A*. Meanwhile, hard joints in *B* are refined by the bottom channel. The two channels of the refinement network correspond to f_{Ar} and f_{Br} respectively, Eq. (2) will be updated as:

$$L_3 = \sum_{\mathbf{x}_i \in A} \mathbf{q}_i \times \mathcal{L}(f_{Ar}(f_b(\mathbf{x}_i)) - \mathbf{y}_i), \tag{3}$$

$$L_4 = \sum_{\mathbf{x}_i \in B} \mathbf{q}_i \times \mathcal{L}(f_{Br}(f_b(\mathbf{x}_i)) - \mathbf{y}_i), \tag{4}$$

where $\mathbf{q}_i$ is set to 1 if the i_{th} joint is a hard joint in *A* or *B*.

Feature fusion: In geometry, 3D pose reconstruction heavily requires its 2D information [18]. The camera pose estimation problem aims to estimate the extrinsic parameters i.e., the rotation matrix *R* and the translation vector **t**, and possibly all or a subset of the intrinsic parameters *K*. *K* is expresses as

$$K = \begin{bmatrix} \alpha f & s & u \\ 0 & f & v \\ 0 & 0 & 1 \end{bmatrix}, \tag{5}$$

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:

$$X = K[R, t]Y. \tag{6}$$

Due to the intermediate supervision of the basic network, the outputs of the basic network are the predicted 3D location features, and they are the input for our refinement network. If the 3D location refinement is only based on the 3D location features from the basic network, the geometrical relationship between 2D and 3D is seriously weakened. In order to fuse 2D advanced features with 3D features, and to reduce the loss of information during the translation from the basic network to the refinement network, we

Table 1
Results on the Human3.6M. Comparison of overall average errors (mm) for different methods using the Human3.6M dataset. '*' represent the method is camera module based method. '–' means that the result of corresponding work is not reported. 'wo dual' indicates the proposed model is implemented using a single-channel model, and 'w dual' for a dual-channel model. 'wo res' means that our model has no residual connection between the basic network and the refinement network.

Methods	Direct	Discuss	Eating	Greet	Phone	Pose	Purchase	Sitting
Ionescu PAMI'14 [4]	132.71	183.55	133.37	164.39	162.12	205.94	150.61	171.31
Li ICCV'15 [8]	–	138.88	96.94	124.74	–	168.08	–	–
Tekin CVPR'16 [15]	102.41	147.72	88.83	125.28	118.02	112.38	129.17	138.89
Zhou* CVPR'16 [22]	87.36	109.31	87.05	103.16	116.18	143.32	106.88	99.78
Moreno CVPR'17 [11]	67.44	63.76	87.15	73.91	**71.48**	69.88	65.08	**71.69**
Martinez ICCV'17 [10]	53.30	60.80	62.90	62.70	86.40	57.80	58.70	81.90
Zhou* PAMI'18 [23]	68.70	74.80	67.80	76.40	76.30	84.00	70.20	88.00
Proposed/w dual/wo res	54.00	61.28	64.35	63.62	87.32	58.74	60.62	83.72
Proposed/wo dual	**52.57**	60.39	61.88	62.25	86.37	**56.99**	**57.74**	80.98
Proposed/w dual	52.83	**59.90**	**61.58**	**61.95**	85.47	57.03	58	81.29

Methods	SittingDown	Smoke	Photo	Wait	Walk	WalkDog	WalkTogether	Average
Ionescu PAMI'14 [4]	151.57	243.03	162.14	170.69	177.13	96.6	127.88	162.1
Li ICCV'15 [8]	–	–	–	–	132.17	69.97	–	–
Tekin CVPR'16 [15]	224.9	118.42	182.73	138.75	55.07	126.29	65.76	124.97
Zhou* CVPR'16 [22]	124.52	199.23	107.42	118.09	114.23	79.39	97.7	112.91
Moreno-Noguer CVPR'17 [11]	98.63	81.33	93.25	74.62	76.51	77.72	74.63	76.47
Martinez ICCV'17 [10]	99.80	69.10	82.40	63.90	50.90	67.10	54.80	67.50
Zhou* PAMI'18 [23]	113.80	78.00	98.40	90.10	62.60	75.10	73.60	79.90
Proposed/w dual/wo res	100.92	69.49	84.97	64.83	51.75	67.92	56.37	68.66
Proposed/wo dual	98.42	68.60	81.42	63.4	49.77	66.79	54.12	66.78
Proposed/w dual	**98.29**	**68.27**	**81.32**	**63.29**	**49.42**	**65.83**	**53.79**	**66.55**

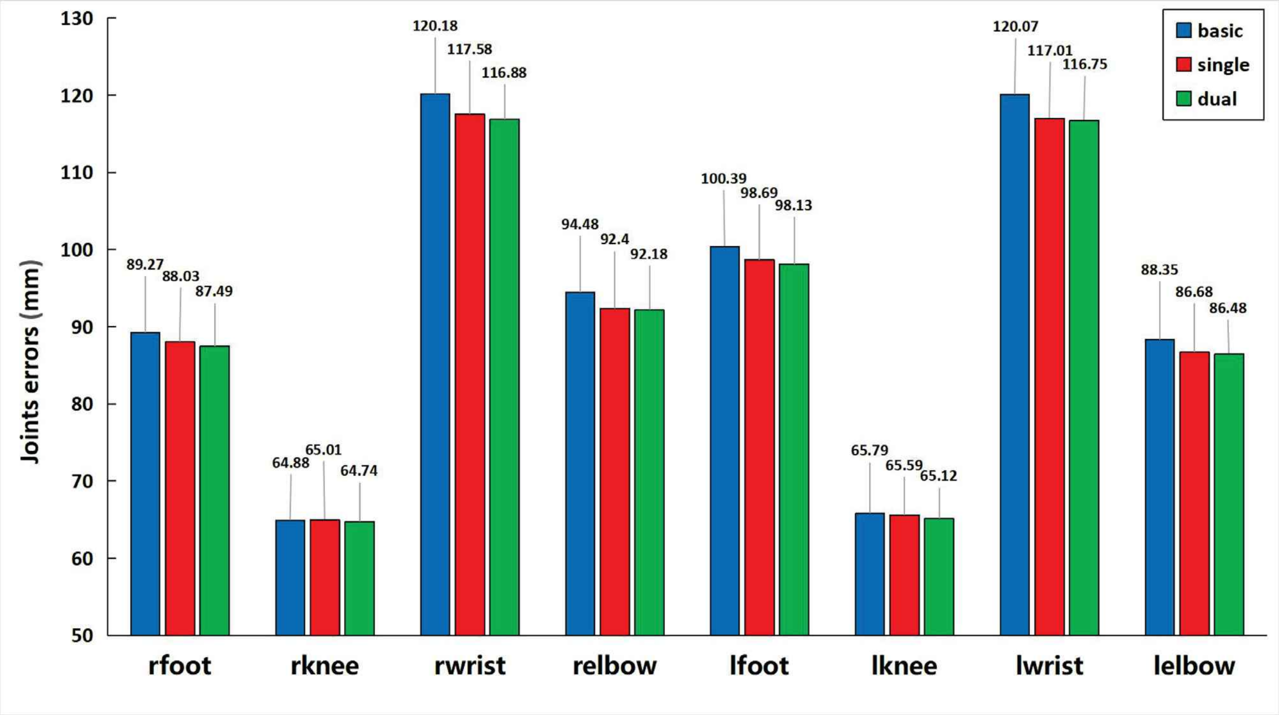


Fig. 5. Errors of hard joints. Average errors (mm) of hard joints of both the basic network and the proposed networks (single-channel model and dual-channel model). 'lelbow' denotes the left elbow of body, 'relbow' for the right one, and others in a similar fashion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

employ a residual connection among the two networks. By doing this, the refinement network can refine hard joints by processing the preliminary 3D locations with further aid of 2D locations evidence. This provides a strong connection between joints without unnecessary interference, and maintains the indispensable geometrical relationship without complex optimization algorithm.

Finally, we formulate our loss function to minimize the prediction error as:

$$L = \min(L_1 + L_3 + L_4). \quad (7)$$

3. Experiments

3.1. Implementation details

We evaluate our proposed approach on a publicly available dataset, namely Human3.6M. Human3.6M is a large scale dataset for 3D human pose sensing, which consists of 3.6 million 3D poses of 11 subjects performing 15 different actions under 4 viewpoints. In this paper, we follow the standard protocol [10], partitioning all

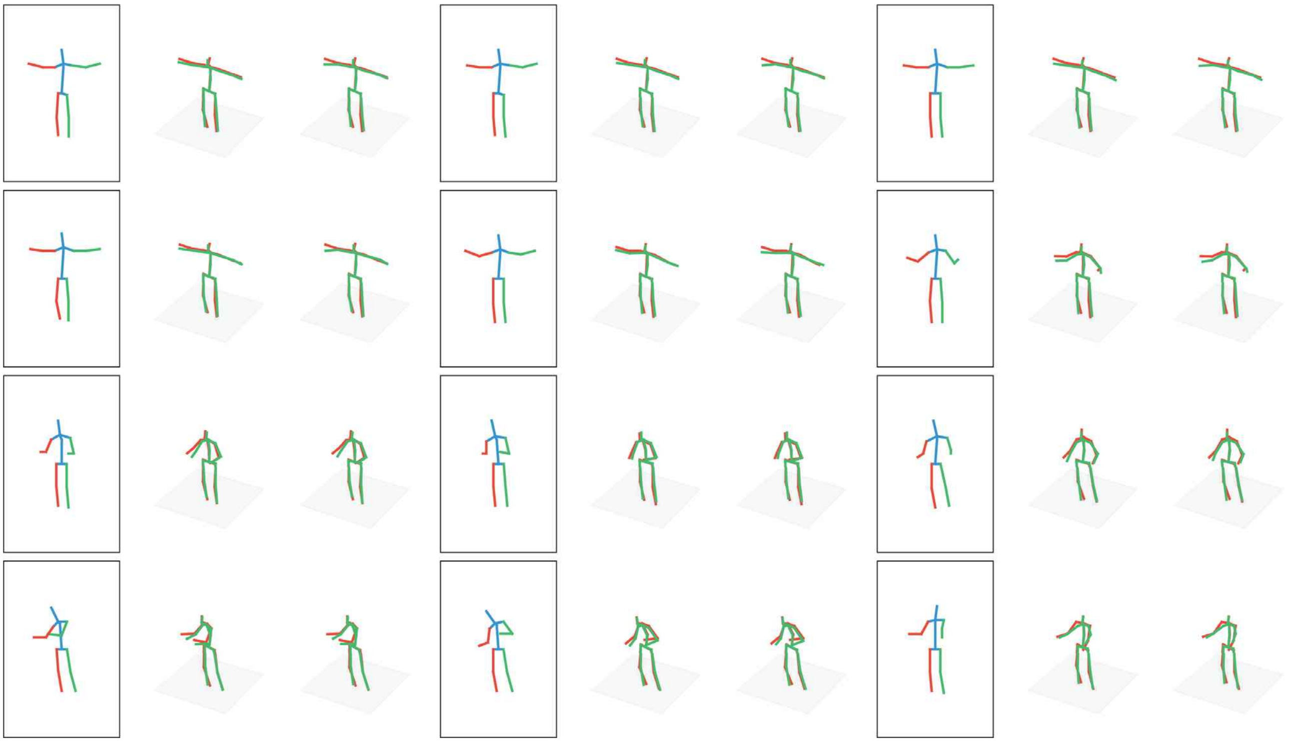


Fig. 6. Examples of selected poses using Human3.6M dataset. The images from left to right in each triplet correspond to the given 2D pose (in a square), overlapping 3D pose pair of the proposed network, and overlapping 3D pose pair of the basic network, respectively. In each overlapping 3D pose pair, the one in red is the ground truth and the other in green is the prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

subjects of dataset for testing ($S1$, $S5$, $S6$, $S7$, $S8$) and training ($S9$, $S11$).

Network architecture: In the proposed approach, the Stacked Hourglass Network [13] is adopted to obtain 2D joints locations, which achieves sufficient accuracy for our 3D-HPE task, unless stated otherwise. The network captures and consolidates information across all scales of image by a repeated bottom-up, top-down structure, and takes full advantage of the good performance of Fully Convolutional Network [9] in 2D human pose estimation (2D-HPE).

The basic network f_b is realized by 2 residual modules, and each residual module contains 2 fully connected layers as shown in Fig. 2. The refinement network is realized by 2 channels, a total of 4 fully connected layers. We empirically set the dimensions of every fully connected layer to 1024, and each layer has Batch Normalization, ReLU activation function and Dropout.

Training details: We use Adam [6] to optimize the network 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, a mini-batch size of 64 for 300 epoches in our work. The learning rate is initialized as 1×10^{-3} and exponential decay.

3.2. Quantitative results

Following previous works, we quantitatively evaluate our method by calculating the Root Mean Square Error (RMSE) between all predicted 3D joint coordinates and the ground truth. The RMSE of our proposed method is compared with those of some related state-of-the-art methods. Although we use a simple fully connected network, the results summarized in Table 1 show that our approach outperforms the state-of-the-art methods. It is probably because that our human structure-aware module makes a distinction between joints. Specially, work [11] explicitly constraints distances between joints by a distance matrix, 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.

As in the comparison with related state-of-the-art methods (in Table 1), the baseline method [10] has the most close errors with our method. Therefore, two statistic tests (T-test and Wilcoxon signed-rank test) are performed between testing errors of our network and the baseline method (in Table 1) to show if the differences are significant or not. Given null hypothesis that the errors of our dual-network in each action are greater than or equal to the baseline [10], we have $P < \alpha(0.05)$ in most of the poses (except for the action of Phoning) for ***T*-test**. Given null hypothesis that the estimation errors of paired samples of our network and the baseline come from the same distribution, we have $P < \alpha(0.05)$ in all actions for the **Wilcoxon signed-rank test**. Therefore, we reject both null hypothesis. The ***T*-test** and the **Wilcoxon signed-rank test** prove that the errors of our network are significantly smaller than those of the baseline (more details are shown in the supplementary file).

To further demonstrate our potential, we show two different structures of our model, which correspond to the single-channel and the dual-channel refinement networks respectively. It is notable that our dual-channel model obtains a relatively slight performance improvement compared with the one obtained by a single-channel, while this is meaningful for 3D-HPE. This feasible idea motivates us to concern the mutual inference between joints in the future work.

In addition, we explore the importance of our feature fusion operation by the residual connection from the basic network to the refinement network. Without involving with the feature fusion operation (i.e without residual connection), the average joint error of our dual-channel network reaches up to **68.66 (mm)**, which is worse than the basic network and our model with residual connection. Refining hard joint coordinates from their 3D location

Table 2
Errors of limb lengths. RMSE (mm) of the bone length for the basic network and the proposed networks (single channel model and dual channel model). '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	Single	Dual
lu_arm	16.90	15.33	14.69
ll_arm	28.29	21.93	20.37
lu_leg	17.21	16.20	15.8
ll_leg	27.15	24.45	24.43
ru_arm	16.45	14.45	14.22
rl_arm	29.99	23.75	21.28
ru_leg	16.62	15.86	15.34
rl_leg	26.98	23.64	23.65
avg	22.45	19.45	18.72

features is likely to be insufficient. Because it enforces to perform location searching in a giant space. After introducing 2D location features, projection clues and constraints between joints could work together, yielding better performance. The experimental results prove this observation.

For more discussion of the effectiveness of our refinement network, the errors of hard joints are separately analyzed and compared, which is shown in Fig. 5. The errors of hard joints are significantly reduced by using our single-channel and dual-channel refinement networks when compared with those of the basic network. With dual-channel network, the errors of wrists reduce 4*mm*, and those of elbows and feet reduce 2 mm. Specifically, we classify the knees as hard joints due to its structural information with feet, while knees have small errors. As such, our single-channel network gets a worse effect at the right knee, and our dual-channel network gets the best results at all hard joints. The improvement demonstrates the effectiveness of our method. Meanwhile, it shows that those hard joints can be further optimized as they still have large errors.

Furthermore, the advantage of our method on reducing the bone length errors is also analyzed, and the results are reported in Table 2. As is depicted, our approach estimates a more reasonable 3D human pose whose limbs length errors reduce significantly by **16%**. Among them, our dual-channel module gets more significant effect on limbs length errors than the single-channel module except lower legs. This is because the purpose of our dual-channel module is to avoid mutual interference of hard joints. In this case, both of the single-channel and dual-channel modules have small errors at knees. The length errors of legs are very close between these two refinement modules.

3.3. Qualitative results

Finally, we show some qualitative results on Human3.6M in Fig. 6. As shown in the first two lines, given 2D poses, the endpoints of estimated 3D poses based on the basic network gravely deviate from the actual coordinates, while our arms extend to a more correct direction. In addition, for the poses in the next two lines, both methods are obviously incorrect in left-wrists, but our poses show more accurate arm lengths.

4. Conclusion

We have proposed a novel human structure-aware network which cascades a dual-channel refinement network with a basic fully connected network for 3D-HPE. The symmetrical structure is designed to avoid the mutual interference of joints with large errors and to promise reliable 3D features. Consideration of the special articulated structure of human body makes a positive contribution to the accurate and reasonable pose learning. Extensive experimental results on Human3.6M have demonstrated the superior performance of the proposed method over various classical and state-of-the-art ones. Also, the experimental results inspire us to pay more attention to hard joints, since they always possess large errors, which is the main challenge in 3D-HPE.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgements

This work was supported in part by:

(i) the National Natural Science Foundation of China (Grant No. 61602313, 61602314);

(ii) Shenzhen Commission of Scientic Research & Innovations under the Grant No. JCYJ20170302153632883;

(iii) Startup Foundation for Advanced Talents, Shenzhen.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2019.05.020.

References

[1] C.-H. Chen, D. Ramanan, 3d human pose estimation= 2d pose estimation+ matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2, 2017, pp. 5759–5767.

[2] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[3] E.P. Ijjina, et al., Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognit. Lett. 83 (2016) 268–277.

[4] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1325–1339.

[5] H. Jiang, 3d human pose reconstruction using millions of exemplars, in: IEEE International Conference on Pattern Recognition, 2010, pp. 1674–1677.

[6] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980 (2014).

[7] J. Lezama, Q. Qiu, P. Musé, G. Sapiro, OL\'E: orthogonal low-rank embedding, a plug and play geometric loss for deep learning, arXiv:1712.01727 (2017).

[8] S. Li, W. Zhang, A.B. Chan, Maximum-margin structured learning with deep networks for 3d human pose estimation, in: IEEE International Conference on Computer Vision, 2015, pp. 2848–2856.

[9] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[10] J. Martinez, R. Hossain, J. Romero, J.J. Little, A simple yet effective baseline for 3d human pose estimation, in: IEEE International Conference on Computer Vision, 206, 2017, pp. 2659–2668.

[11] F. Moreno-Noguer, 3d human pose estimation from a single image via distance matrix regression, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 1561–1570.

[12] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: International Conference on Machine Learning, 2010, pp. 807–814.

[13] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision, 2016, pp. 483–499.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[15] B. Tekin, A. Rozantsev, V. Lepetit, P. Fua, Direct prediction of 3d body poses from motion compensated sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 991–1000.

[16] M. Tong, Y. Liu, T.S. Huang, 3d human model and joint parameter estimation from monocular image, Pattern Recognit. Lett. 28 (7) (2007) 797–805.

[17] C. Wang, Y. Wang, Z. Lin, A.L. Yuille, W. Gao, Robust estimation of 3d human poses from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2361–2368.

[18] G. Wang, Q.J. Wu, Simplified camera projection models, in: Guide to Three Dimensional Structure and Motion Factorization, 2011, pp. 29–41.

[19] J. Yan, S. Shen, Y. Li, Y. Liu, An optimization based framework for human pose estimation, IEEE Signal Process. Lett. 17 (8) (2010) 766–769.

[20] X. Zhou, S. Leonardos, X. Hu, K. Daniilidis, et al., 3d shape estimation from 2d landmarks: a convex relaxation approach., in: IEEE Conference on Computer Vision and Pattern Recognition, 2, 2015, pp. 4447–4455.

[21] X. Zhou, M. Zhu, S. Leonardos, K. Daniilidis, Sparse representation for 3d shape estimation: a convex relaxation approach, IEEE Trans. Pattern Anal. Mach. Intell. 39 (8) (2017) 1648–1661.

[22] X. Zhou, M. Zhu, S. Leonardos, K.G. Derpanis, K. Daniilidis, Sparseness meets deepness: 3d human pose estimation from monocular video, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4966–4975.

[23] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K.G. Derpanis, K. Daniilidis, Monocap: monocular human motion capture using a cnn coupled with a geometric prior, IEEE Trans. Pattern Anal. Mach. Intell. (2018).