Coarse-to-Fine 3D Human Pose Estimation

Yu Guo, Lin Zhao, Shanshan Zhang(✉), and Jian Yang(✉)

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{csyguo,linzhao,shanshan.zhang,csjyang}@njust.edu.cn

Abstract. Leveraging powerful deep convolutional networks, 2d human pose estimation has achieved great success. On the other hand, 3d human pose estimation is still a challenging task that attracts great attention. Due to the inherent depth ambiguity in 2d to 3d mapping, conventional methods are typically not able to predict 3d locations precisely, especially for the joints far from the torso. In this paper, we propose a coarse-to-fine model to predict 3d joint locations progressively. We observe that some joints like shoulders and hips are relatively easy to get precise 3d locations, which can be utilized to facilitate the prediction of hard joints that are far from the torso. To make this happen, a set of constraints based on human limb length ratio prior is proposed to guide the model to generate reasonable predictions. We conduct experiments on the Human3.6M dataset. Comparison of experimental results on the benchmark dataset turns out that our approach outperforms the baseline method.

Keywords: 3D human pose estimation · Human limb length ratio prior · Deep learning

1 Introduction

Human pose estimation, also called as human keypoints detection, has received extensive attention in recent years. The primary purpose of human pose estimation is to predict human joint locations from monocular RGB information. Human pose estimation is a classical middle-level computer vision task and can greatly facilitate other related high-level tasks such as pedestrian detection [28] and action recognition [7].

Following the success of deep convolutional networks, current 2d human pose estimation methods perform well even in complex outdoor environments. Figure 1 shows typical 2d human pose estimation results predicted by stacked hourglass [18] on Human3.6M dataset [11]. However, unlike on Human3.6M dataset [11]. However, unlike 2d human pose estimation, it is challenging to obtain annotated data for 3d human pose estimation tasks. Most 3d human

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11903, pp. 579–592, 2019.
https://doi.org/10.1007/978-3-030-34113-8_48

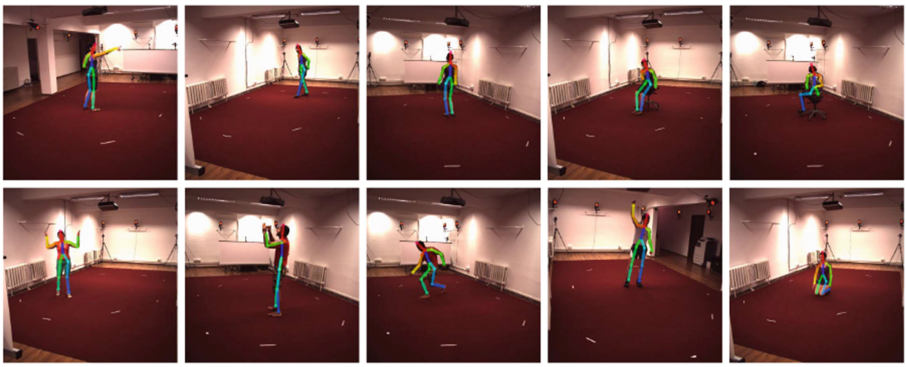

Fig. 1. Typical 2d human pose estimation results produced by stacked hourglass model [18]. Images are from Human3.6M dataset. We can see that stacked hourglass model performs well on Human3.6M dataset.

pose datasets only contain indoor data collected in a laboratory environment, which leads to lack of diversity. Thus, models tend to overfit when training on such datasets. Besides, ambiguity is a widespread problem when mapping 2d to 3d, which also results in unreasonable predictions.

In this paper, we propose a novel coarse-to-fine method for 3d human pose estimation. From our analysis, we find that current models usually produce large errors when predicting keypoints located at the end of limbs, such as wrists and ankles. In contrast, joints like shoulders and hips are relatively easy to predict. Table 1 shows detailed statistics about errors of each joint by [14]. We assume that *easy* joints can be helpful to guide the prediction of *hard* joints. Therefore we propose a coarse-to-fine method to predict different joints in a progressive way. An intuitive way to deal with ambiguity in 3d human pose estimation is to leverage the prior of human structure. For instance, Dabral *et al.* [5] use legal angular constraints in their model. Here, we propose a set of limb length ratio (LLR) constraints to reduce the shifts of joints from the true locations.

Our contributions can be summarized as follows:

- We propose a specific coarse-to-fine method for 3d human pose estimation task to enhance precision of the joints far from the torso. Based on the statistical analysis of predictions produced by the previous state-of-the-art method, we divide joints into three groups according to different difficulty levels. *Easy* joints are predicted first, and then they are used to facilitate the prediction of harder joints.
- A set of human limb length ratio (LLR) constraints based on the statistics of physical human body structure are used to avoid unreasonable predictions, allowing the model to perform more robust on *hard* joints.
- By combining the coarse-to-fine model and LLR constraints, our method outperforms the baseline on the Human 3.6M dataset. Especially the improvement is more significant for those joints far from the torso.

Table 1. Detailed statistics on the error of each joint produced by [14]. Numbers denote the error of each joint in millimeters. Under protocol 2, the model predictions are post-processed with rigid alignments.

Joint	Hip	RHip	RKnee	RFoot	LHip	LKnee	LFoot	Spine	Thorax	Neck
Protocol 1	0.00	23.28	67.74	92.72	23.28	67.76	102.5	44.94	50.58	62.89
Protocol 2	33.48	43.67	53.09	71.12	38.56	54.23	77.06	34.03	27.55	37.37
Joint	Head	LShoulder	LElbow	LWrist	RShoulder	RElbow	RWrist			
Protocol 1	73.32	64.15	88.17	120.38	66.35	94.69	120.95			
Protocol 2	44.21	43.82	58.10	90.76	37.08	63.12	90.14			

2 Related Work

Since our method is specifically designed for the 3d human pose estimation task, we will first review recent works on it. Moreover, we will review recent works on the usage of human structure prior to the task for human pose estimation.

2.1 3D Human Pose Estimation

The topic of 3d human pose estimation attracts increasing attention in recent years due to its potentially broad application prospects. The purpose of 3d human pose estimation task is to estimate accurate spatial position coordinates of human keypoints from RGB images. It is proven that positions of human keypoints are beneficial for generic action recognition tasks in previous works [13,22]. In the current stage, it is almost impossible to predict 3d coordinates in the world coordinate system, as is declared in [14]. Thus most of the current methods predict coordinates in the camera coordinate system [5,9,25]. In this paper, our model predicts 3d human keypoint locations in the camera coordinate system as well.

Various types of methods, as well as diverse representations are proposed for 3d human pose estimation. A typical way of 3d human pose estimation is to use 3d coordinates to represent human keypoint locations and to regress coordinates from a single RGB image directly, as is proposed in [21]. However, the mapping from RGB images to 3d coordinates is so complex that it is challenging to learn the potential knowledge between images and coordinates. In order to overcome this problem, volumetric representation is used as supervision [21,27], which contains richer information than coordinates. Volumetric representation, however, leads to a huge number of model parameters and increasing computational complexity. A compromise solution is to use 3d coordinates as supervision, leveraging 2d human pose predictions at the same time. With the help of powerful convolutional neural networks (CNN), the performance of 2d human pose estimation has great improvements in recent years. A simple yet effective method is to use 2d human pose predictions as input to regress 3d coordinates of human keypoints [14]. Based on this work, [9] combines temporal information with 2d to 3d pose regression, which allows the model to perform well. However, temporal

information puts high demands on the data, and also such a model costs too much computation, making it hard to be used in practical applications.

These works make good progress, but it is worth mentioning that the points far from torso flutter heavily in their predictions. This phenomenon is consistent with the problem in 2d pose estimation, as proposed in [24]. In this paper, we propose a coarse-to-fine method, which takes 2d human pose prediction from a single image as input and predicts the 3d coordinates of human keypoints. We divide human keypoints to three groups according to different difficulty levels. The further the keypoints are from the human torso, the harder they are for a model to predict. Our model predicts *easy* keypoints first and then predicts *medium* and *hard* keypoints in turn, leveraging former prediction results.

2.2 Human Structure Prior in Pose Estimation

In previous works, models often generate unreasonable predictions, which makes human structure prior indispensable in human pose estimation tasks. In 2d human pose estimation, [4] leverages generative adversarial networks to guide a model to learn human structure prior implicitly. [5] proposes angular constraints based on the human prior that the range of motions of human joints is limited and symmetry. These constraints are reasonable while the limb length ratio can be another useful constraint, whose distribution is proven to obey specific rules [6]. In this paper, we propose a set of constraints based on the human limb length ratio, and experiments demonstrate it is helpful for a model to get better performance in the task of 3d human pose estimation.

3 Method

In this section, we will discuss the method proposed for 3d human pose estimation. We start with the coarse-to-fine method and introduce the limb length ratio (LLR) constraint to solve the problem better.

3.1 Coarse-to-Fine Model

In previous works, models usually perform worse when predicting keypoints far from the torso such as wrists and ankles. In order to overcome this problem, we propose a coarse-to-fine method. In our method, we first divide keypoints into three groups according to the prediction difficulty. From Table 1, we can observe that the closer the keypoints are to the body torso, the more accurate the model prediction is. For instance, the model performs better when predicting the location of the head than elbows; and performs worse when predicting ankles than knees. Thus we can divide keypoints, according to their distance to the torso, into three groups: *easy*, *medium*, and *hard*. A detailed demonstration is shown in Fig. 2. According to Table 1, we classify head, spine, thorax, hip and shoulder as *easy* joints, elbow and knee as *medium* joints, wrists and ankles as *hard* joints, as shown in Fig. 2.

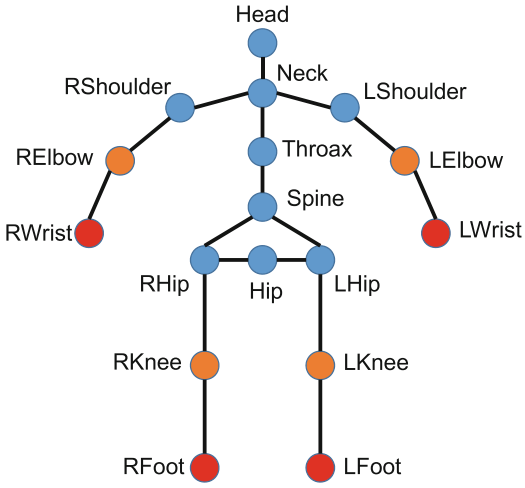

Fig. 2. Keypoints grouped by prediction difficulties. Circles colored in blue, orange and red denote *easy*, *medium* and *hard* joints respectively. The position of hip is the midpoint of left hip and right hip. (Color figure online)

Based on the characteristic of different difficulty levels of joints, we design a specific coarse-to-fine model. The network structure of our model is shown in Fig. 3. The input of our model is 2d keypoints predictions produced by a 2d human pose estimator, and the output is predictions of 3d human keypoints coordinates. As we can see in Fig. 3, our model contains three stages. In the first stage, we predict *easy* joints by using a simple fully-connected network, which is effective in a regression task mapping 2d coordinates to 3d coordinates [14]. In the second and third stage, we predict *medium* and *hard* keypoints, taking both 2d keypoints and 3d coordinates predictions produced in the previous stage(s) as input. Therefore we can leverage predicted 3d joint coordinates as auxiliary information to guide the model to produce more accurate predictions for challenging keypoints. In order to merge 2d keypoints and 3d keypoint predictions produced in previous stages, we adopt channel wise self-attention blocks, as is proposed in [10], to guide the model to assign appropriate weights for predicted 3d keypoint coordinates in the second and third stages. We compute Euclidean distance between 3d keypoints prediction and groundtruth as the keypoints loss L_K,

$$L_K(x, y) = \frac{1}{m} \sum_{i=1}^{m} \|x_i - y_i\|, \tag{1}$$

where x, y stands for the model prediction and groundtruth respectively, m stands for the number of keypoints. Considering that our model produces predictions in three stages, the loss function is written as

$$L_{CTF}(x, y) = \theta_1 L_K(x_e, y_e) + \theta_2 L_K(x_m, y_m) + \theta_3 L_K(x_h, y_h), \tag{2}$$

where subscript e, m, h denotes *easy*, *medium*, and *hard* keypoints respectively.

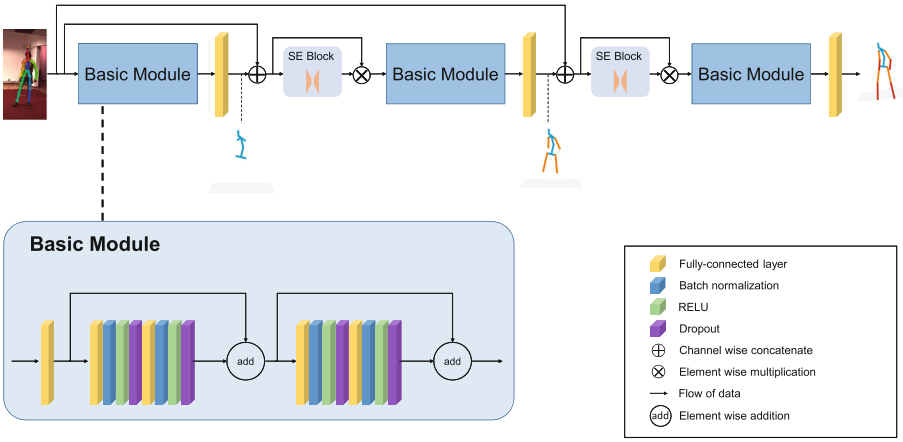

Fig. 3. Network structure of our coarse-to-fine model. For a given RGB image, we first obtain 2d keypoint locations via a 2d human pose estimator. Then we design the coarse-to-fine model in order to predict 3d keypoint coordinates from 2d keypoints. Our method can be divided into 3 stages and we predict positions of *easy*, *medium* and *hard* joints in order. During the second and the third stages, the model leverages predictions from previous stage(s).

3.2 LLR Constraint

Human pose prior knowledge is helpful in the 3d human pose estimation task; and human limb length ratio (LLR) is an important prior, which is studied in [6]. Within the best of our knowledge, few researches focus on LLR prior, which helps predict accurate 3d coordinates. In this paper, we propose a set of LLR constraints based on the LLR prior. According to the research of *De Leva* [6], we can assume that the distribution of adult limb length ratio obeys normalization distribution. Therefore we can census the dataset to get the mean value and stand deviation of the limb length ratio of the dataset.

The length of a limb can be computed as follows,

$$l(x_1, x_2) = \|x_1 - x_2\|, \tag{3}$$

where x_1, x_2 stands for 3d coordinates of corresponding keypoints lying at the ends of limbs. The limb length ratio between limb p and limb q can be computed as follows,

$$r(p, q) = \frac{l(p_{x_1}, p_{x_2})}{l(q_{x_1}, q_{x_2})}, \tag{4}$$

where p_{x_1}, p_{x_2}, q_{x_1} and q_{x_2} stand for 3d keypoint coordinates lying at the ends of limb p and limb q respectively. Then the LLR loss can be written as

$$L_{LLR}(X) = \frac{1}{m}\sum_{i=1}^{m}\left(1 - \frac{1}{\sqrt{2\pi}s}exp(-\frac{1}{2}\frac{\left(r(X_{i_p}, X_{i_q}) - \overline{R}\right)^2}{s}\right), \tag{5}$$

Table 2. Comparison to current state-of-the-art methods on the Human3.6M validation set under protocol 1. Bold indicates the best results.

Protocol 1	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
LinKDE [11]	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Tekin *et al.* [26]	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou *et al.* [30]	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Du *et al.* [8]	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Park *et al.* [20]	100.3	116.2	90.0	116.5	115.3	149.5	117.6	106.9	137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3
Zhou *et al.* [31]	91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Nie *et al.* [19]	90.1	88.2	85.7	95.6	103.9	103.0	92.4	90.4	117.9	136.4	98.5	94.4	90.6	86.0	89.5	97.5
Mehta *et al.* [15]	57.5	68.6	59.6	67.3	78.1	82.4	56.9	69.1	100.0	117.5	69.4	68.0	76.5	55.2	61.4	72.9
Mehta *et al.* [16]	62.6	78.1	63.s4	72.5	88.3	93.8	63.1	74.8	106.6	138.7	78.8	73.9	82.0	55.8	59.6	80.5
Martinez *et al.* [14]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
CTF (ours)	49.9	55.3	56.7	57.4	66.7	77.1	53.2	55.5	71.9	89.4	60.2	58.2	62.9	48.8	51.2	61.2
CTF+LLR (ours)	**49.4**	**54.3**	**55.7**	**56.9**	**66.4**	**74.5**	**53.2**	**55.4**	**71.7**	**89.0**	**60.0**	**57.0**	**62.7**	**48.0**	**50.7**	**60.6**

Table 3. Comparison to current state-of-the-art methods on the Human3.6M validation set under protocol 2. Bold indicates the best results.

Protocol 2	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Akhter and Black [1]	199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3	160.7	173.7	177.8	181.9	176.2	198.6	192.7	181.1
Ramakrishna *et al.* [23]	137.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1	168.6	175.6	160.4	161.7	150.0	174.8	150.2	157.3
Zhou *et al.* [29]	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1	137.5	106.0	102.2	106.5	110.4	115.2	106.7
Bogo *et al.* [3]	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	86.8	79.7	87.7	82.3
Moreno-Noguer [17]	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	71.5	78.0	73.2	74.0
Martinez *et al.* [14]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
CTF+LLR (ours)	**38.8**	**42.1**	**44.4**	**46.0**	**49.8**	**53.4**	**40.5**	**39.3**	**54.5**	**63.5**	**47.6**	**43.2**	**48.6**	**36.6**	**41.5**	**46.5**

where X_{i_p} and X_{i_q} denote the limb in the ratio pair respectively, $\overline{R}$ and s denote the mean value and standard deviation of the limb length ratio of a chosen pair $r(X_{i_p}, X_{i_q})$ that computed on the training set respectively. In addition, we use the Gaussian function to punish the ratio offset. Then the final loss function is

$$Loss = \alpha L_{CTF} + \beta L_{LLR}, \tag{6}$$

where α and β are hyper-parameters and denote scale coefficients of the corresponding loss items.

4 Experiments

In this section, we will first describe the implementation details, followed by experimental results on the Human3.6M dataset [11]. In addition, intuitive comparisons between our model and benchmark methods are present.

Table 4. Comparison of the baseline and our method w.r.t the prediction errors of *medium* and *hard* keypoints.

Joints		Protocol 1			Protocol 2		
		Baseline [18]	Ours	Δ	Baseline [18]	Ours	Δ
Medium	LKnee	67.8	59.2	−8.6	54.2	48.3	−5.9
	RKnee	67.7	60.4	−7.3	53.1	48.5	−4.6
	LElbow	88.2	79.2	−9.0	58.1	52.2	−5.9
	RElbow	94.7	83.9	−10.8	63.1	55.6	−7.5
Hard	LFoot	102.5	89.9	−12.6	77.1	65.6	−11.5
	RFoot	92.7	81.9	−10.8	71.1	61.4	−9.8
	LWrist	120.4	105.1	−15.3	90.8	79.0	−11.8
	RWrist	121.0	102.7	−18.2	90.1	75.7	−14.4

4.1 Dataset

We conduct experiments on the Human3.6M dataset to demonstrate the performance of our method. Human3.6M is a widely used dataset in the field of 3d human pose estimation, which contains comprehensive annotations. The data of Human3.6M dataset are collected in a laboratory environment, including 11 professional actors and 17 scenarios. 3d human keypoint position annotations are obtained from a high-speed motion capture system with 4 calibrated cameras. In this paper, we choose 5 actors as the training set and 2 actors as the validation set, which is consistent with widely used protocols [12,14,27]. It is worth mentioning that we do not leverage the temporal information considering real-time performance.

4.2 Implementation Details

In our coarse-to-fine method, we use the predictions of stacked hourglass [18], a state-of-the-art 2d human pose estimation method, as the input of our coarse-to-fine method. A prediction of stacked hourglass includes 16 keypoints. We reshape each 2d human pose prediction to a vector with shape 1×32 and reshape corresponding 3d human pose ground truth to a vector with shape 1×48 during data preprocessing. The 3d human pose ground truth coordinate is transformed to the camera coordinate system. In order to facilitate comparisons with other methods, we set the keypoint *Hip* as the coordinate system origin, which is the midpoint of the left hip and right hip, following [9,14]. In order to make the model easier to convergence, we normalize 2d pose predictions and 3d pose ground truth with mean and variance calculated in the training set. In order to avoid the gradient explosion problem, we clip the maximum L2 norm of gradient every time backpropagation is operated. The model is trained with 128 batch size and 1.22 million iterations in total; the initial learning rate is set to 1×10^{-3}, which is decreased by 0.96 every 10k iterations.

All experiments are conducted on one Nvidia Tesla K80 GPU with 12 Gigabyte memory.

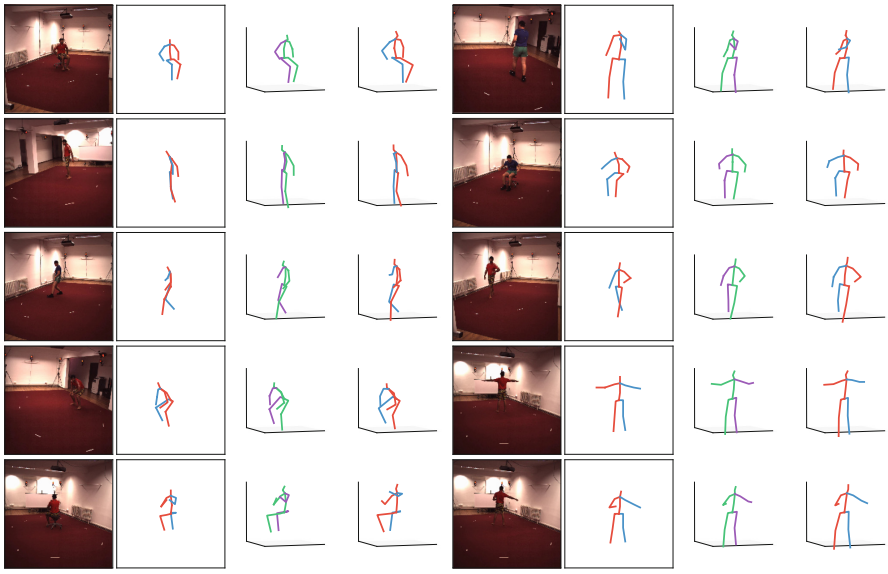

Fig. 4. Qualitative results of our method on the Human3.6M dataset. Each row of the figure contains 2 samples and each sample contains 4 columns. In each sample, each column represents RGB image, 2d human pose prediction produced by stacked hourglass model [18], 3d human pose prediction of our method and the ground truth of 3d human pose in turn. In order to more clearly present the 3d predictions, we rotate the figures in the third and forth columns slightly around the y axis.

4.3 Comparison with State-of-the-Art Methods

In Table 2, we present the results of our methods and make a comparison with the state-of-the-art methods under protocol 1. We can see clearly that our coarse-to-fine method performs well on Human3.6M dataset. When combined with LLR loss, the performance of our method is further improved and decreases the average error to 60.6 mm. Under protocol 2, rigid alignment is applied to the predictions and our method outperforms comparison methods on every action, as shown in Table 3. In Table 4, we can clearly see that our method, which combines LLR loss and coarse-to-fine method, outperforms the baseline method when predicting *medium* and *hard* keypoints. Figure 4 presents some examples of our predicted 3d human poses on the Human3.6M dataset.

In order to explore the generalization performance of our method, we conduct qualitative experiments on MPII dataset [2] and make a comparison between our method and [14], as is shown in Fig. 5. We can see that in most situations, our method produces more reasonable predictions compared with [14] even in wild scenes. While it is worth mentioning that the occlusion of 2d joints has a huge negative impact on 3d prediction.

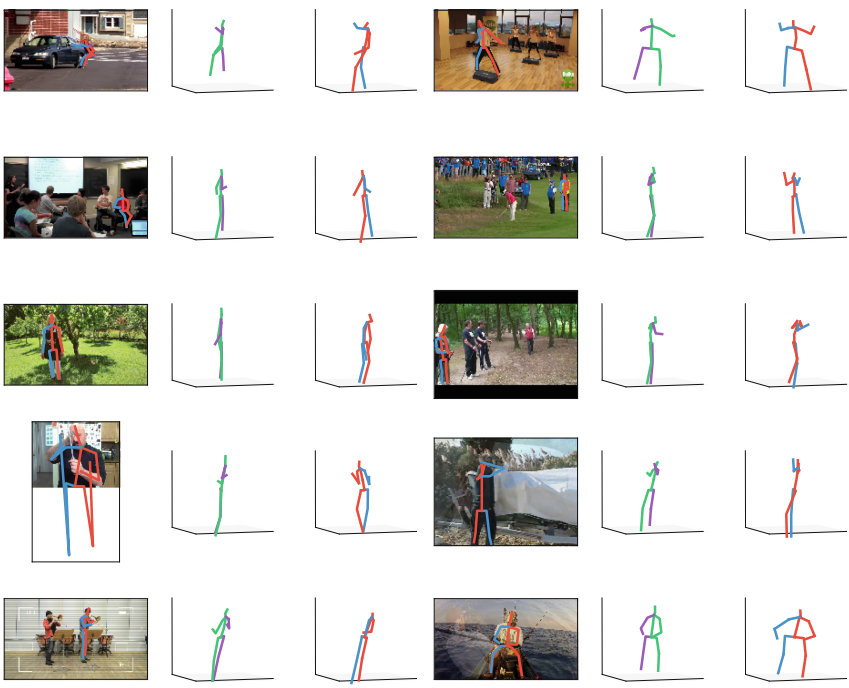

Fig. 5. Qualitative results on the MPII dataset [2]. Each row contains two samples and each sample includes three columns. In each sample, each column represents the RGB image with corresponding 2d human pose prediction, 3d predictions of [14] and 3d predictions of our method in turn.

5 Conclusion

In this paper, we propose a coarse-to-fine method for 3d human pose estimation and a set of human structure based limb length ratio constraints. Experimental results indicate that our method is useful, mainly when predicting challenging keypoints that are far from the torso. Encouraged by the current results, we will investigate how to explore context information to improve the performance further.

Acknowledgements. The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Natural Science Foundation of China under Grant No. U1713208, 61702262 and 61802189, Funds for International Cooperation and Exchange of the National Natural Science Foundation of China under Grant No. 61861136011, Natural Science Foundation of Jiangsu Province, China under Grant No. BK20181299 and BK20180464, the Fundamental Research Funds for the Central Universities under Grant No. 30918011322 and 30918014107, Program for Changjiang Scholars, CCF-Tencent Open Fund No. RAGR20180113, and Young Elite Scientists Sponsorship Program by CAST No. 2018QNRC001.

References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR, pp. 1446–1455 (2015)
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: CVPR (2014)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 561–578. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_34
4. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial PoseNet: a structure-aware convolutional network for human pose estimation. In: ICCV, pp. 1212–1221 (2017)
5. Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A., Jain, A.: Learning 3D human pose from structure and motion. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 679–696. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_41
6. De Leva, P.: Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters. J. Biomech. **29**(9), 1223–1230 (1996)
7. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR, pp. 1110–1118 (2015)
8. Du, Y., et al.: Marker-less 3D human motion capture with monocular image sequence and height-maps. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_2
9. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3D human pose estimation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 69–86. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_5

10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR, pp. 7122–7131 (2018)
13. Luvizon, D.C., Picard, D., Tabia, H.: 2D/3D pose estimation and action recognition using multitask deep learning. In: CVPR, pp. 5137–5146 (2018)
14. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: ICCV, pp. 2640–2649 (2017)
15. Mehta, D., et al.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 3DV, pp. 506–516 (2017)
16. Mehta, D., et al.: VNect: Real-time 3D human pose estimation with a single RGB camera. ACM Trans. Graph. **36**(4), 44 (2017)
17. Moreno-Noguer, F.: 3D human pose estimation from a single image via distance matrix regression. In: CVPR, pp. 2823–2832 (2017)
18. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
19. Nie, B.X., Wei, P., Zhu, S.C.: Monocular 3D human pose estimation by predicting depth on joints. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3467–3475. IEEE (2017)
20. Park, S., Hwang, J., Kwak, N.: 3D human pose estimation using convolutional neural networks with 2D pose information. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 156–169. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_15
21. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR, pp. 7025–7034 (2017)
22. Popa, A.I., Zanfir, M., Sminchisescu, C.: Deep multitask architecture for integrated 2D and 3D human sensing. In: CVPR, pp. 6289–6298 (2017)
23. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3D human pose from 2D image landmarks. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 573–586. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_41
24. Ronchi, M.R., Perona, P.: Benchmarking and error diagnosis in multi-instance pose estimation. In: ICCV, pp. 369–378 (2017)
25. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 536–553. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_33
26. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct prediction of 3D body poses from motion compensated sequences. In: CVPR, pp. 991–1000 (2016)
27. Trumble, M., Gilbert, A., Hilton, A., Collomosse, J.: Deep autoencoder for combined human pose estimation and body model upscaling. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 800–816. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_48
28. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in CNNS. In: CVPR, pp. 6995–7003 (2018)

29. Zhou, X., Zhu, M., Leonardos, S., Daniilidis, K.: Sparse representation for 3D shape estimation: a convex relaxation approach. IEEE Trans. Pattern Anal. Mach. Intell. **39**(8), 1648–1661 (2017)
30. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3D human pose estimation from monocular video. In: CVPR, pp. 4966–4975 (2016)
31. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 186–201. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_17