Orthographic Projection Linear Regression for Single Image 3D Human Pose Estimation

Yahui Zhang

Shaodi You

Theo Gevers

UNIVERSITEIT VAN AMSTERDAM

Introduction

Motivation

Current 3D human pose datasets are collected in indoor environment, limiting the generalization of learning-based approaches for 3D human pose estimation.

- Challenge
  - 2D in-the-wild images are extremely complex.
  - In-the-wild images do not have corresponding 3D ground truth.

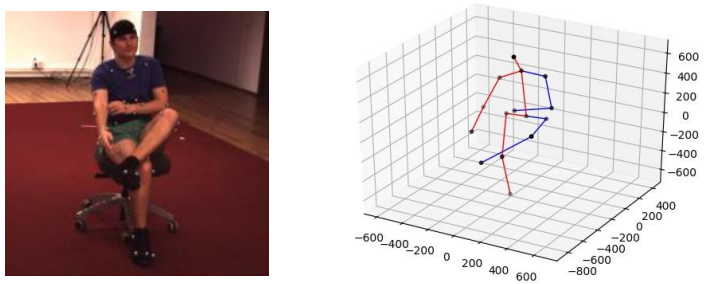

Fig. 1. Indoor image and corresponding 3D ground truth[1]

No
3D labels

Fig. 2. In-the-wild image with no 3D ground truth[2]

1. C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments,” TPAMI, 2013.
2. M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in CVPR, 2014.

Introduction

Goal

The goal of this paper is to regress 3D human joint locations in camera coordinates from a single image.

- Solution

  Propose an orthographic projection linear regression module.

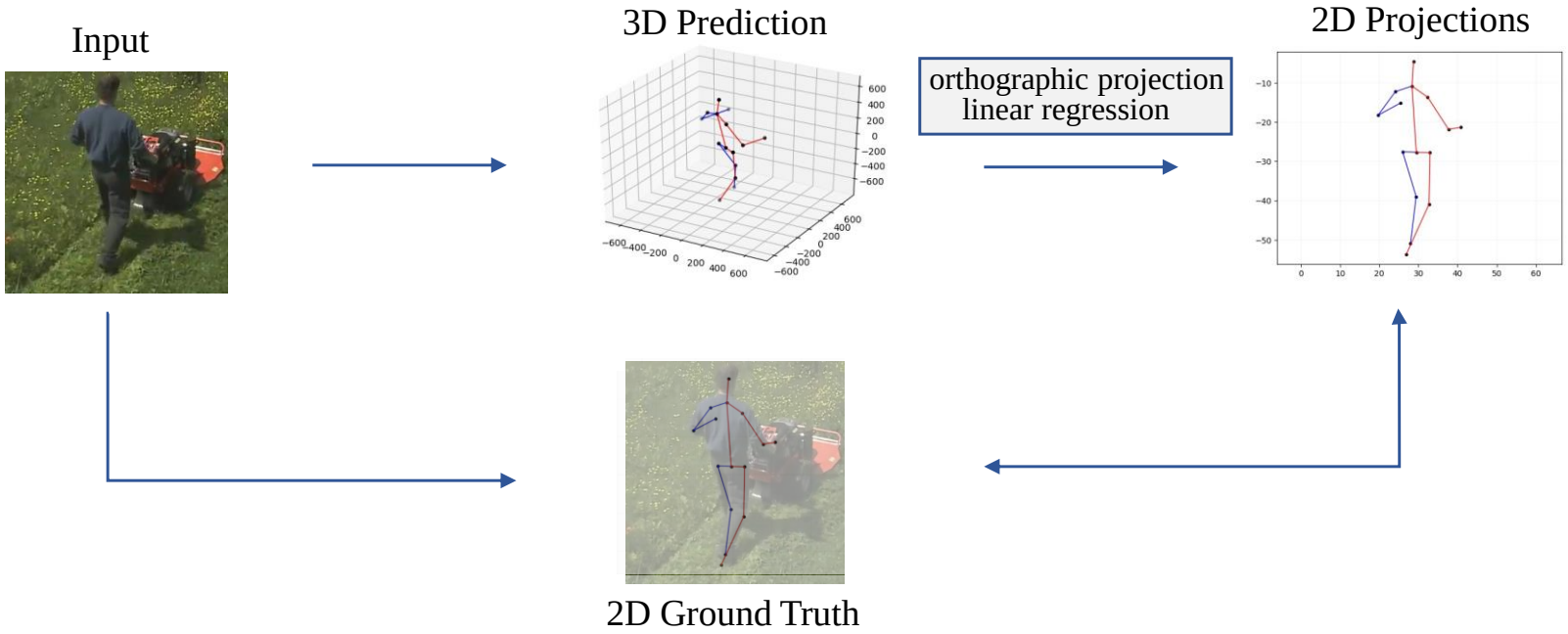

Fig. 3. General procedure of our method to connect in-the-wild images and 3D predictions.

Our Approach

Definition

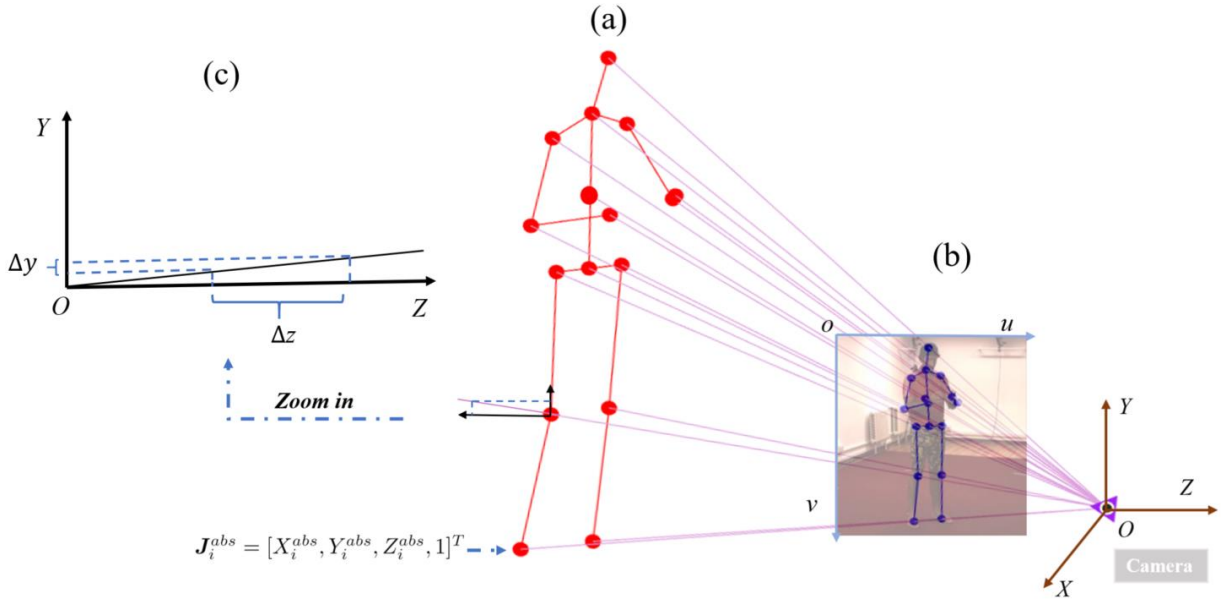

Fig. 4. Perspective projection from (a) the 3D pose to (b) the 2D pose with illustration of small angle problem (c).

- Human pose representations: a set of joints, $\mathbf{P}_{3D}^{abs} = [\boldsymbol{J}_1^{abs}, \boldsymbol{J}_2^{abs}, ..., \boldsymbol{J}_n^{abs}]$, where $\boldsymbol{J}_i^{abs} = [X_i^{abs}, Y_i^{abs}, Z_i^{abs}, 1]^T$
- 2D projections $\mathbf{p}_{2D}$, a 3 by *n* matrix with $\boldsymbol{j}_i^{abs} = [x_i^{abs}, y_i^{abs}, 1]^T$

Camera Model

Given intrinsic ***(K)*** and extrinsic ***(R*** and ***T)*** parameters, 2D projections are obtained by:

$$\mathbf{p}_{2D} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3D}^{abs} \quad (1)$$

Small angle problem arises:
resulting in overfitting in the depth dimension.

Our Approach

Orthographic Projection Linear Regression

***Step 1*: Orthographic Projection.**

$$\mathbf{p}_{2D} = \Pi \mathbf{P}_{3D},$$

$$\Pi = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad (2)$$

where $\mathbf{P}_{3D}$ is root-relative 3D joint locations.

***Step 2*: Constrained Linear Regression.**

$$\mathbf{p}_{2D} = [\boldsymbol{S}|\boldsymbol{t}]\Pi \mathbf{P}_{3D}, \quad (3)$$

where $\boldsymbol{S}$ and $\boldsymbol{t}$ indicate scale and translation parameters.

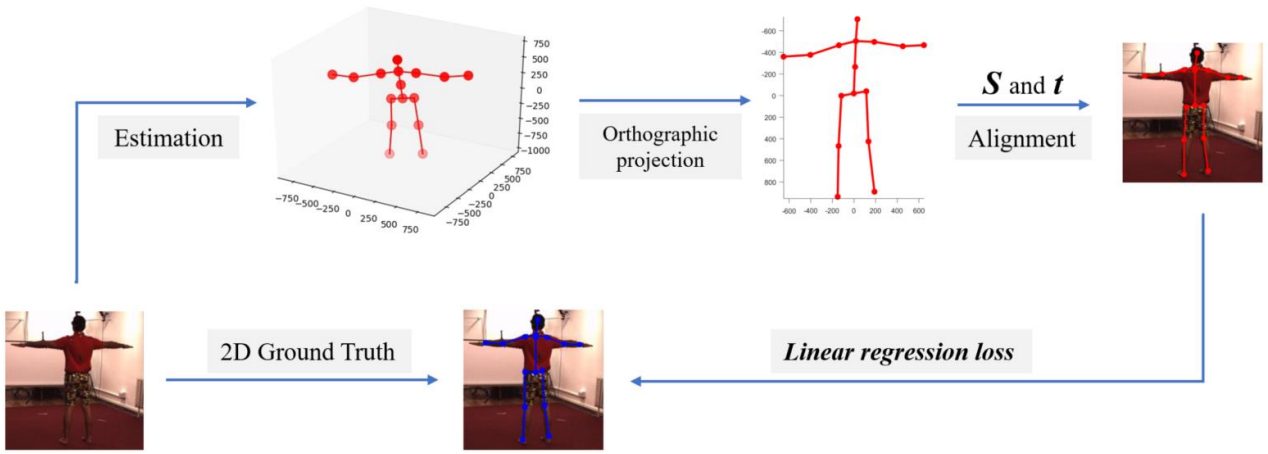

Fig. 5. The general idea of matching 3D with 2D poses by the orthographic projection linear regression method.

➢ Optimization

The linear regression computes the scaling and translation by minimizing:

$$\arg\min_{\boldsymbol{S},\boldsymbol{t}} \|[\boldsymbol{S}|\boldsymbol{t}]\Pi \mathbf{P}_{3D} - \mathbf{p}_{2D}\|_2^2. \qquad (4)$$

Our Approach

Architecture

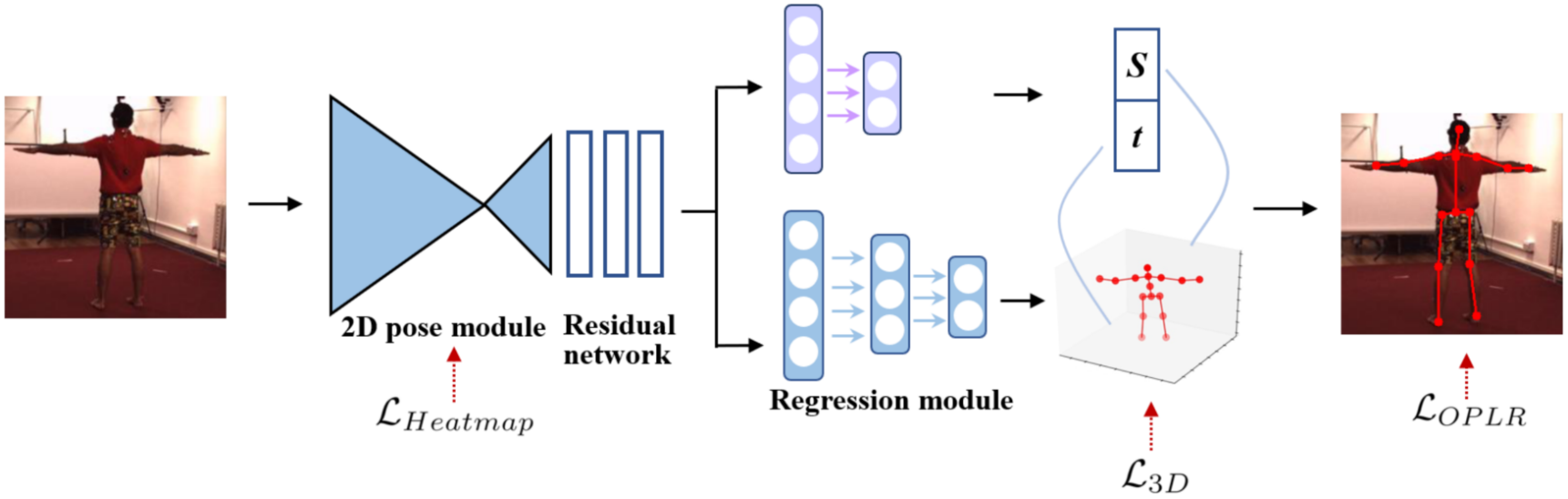

Fig. 6. The overview of the proposed framework.

Loss Function

$$\mathcal{L}_{pose} = \lambda_{hm}\mathcal{L}_{Heatmap} + \mathcal{L}_{3D} + \lambda_{OPLR}\mathcal{L}_{OPLR} \quad (5)$$

Specifically, $\mathcal{L}_{Heatmap} = \left\|\mathbf{HM} - \mathbf{HM}^{\mathrm{GT}}\right\|_2$, $\mathcal{L}_{3D} = \left\|\mathbf{P}_{\mathrm{3D}} - \mathbf{P}_{\mathrm{3D}}^{\mathrm{GT}}\right\|_2$, $\mathcal{L}_{OPLR} = \left\|[\boldsymbol{S}|\boldsymbol{t}]\Pi\mathbf{P}_{\mathrm{3D}} - \mathbf{p}_{\mathrm{2D}}^{\mathrm{GT}}\right\|_2$

where **HM** denotes heatmap.

Experiments

Datasets

Current public datasets: Human3.6m [1] and MPI-INF-3DHP[2]

- Metric
  - Human3.6m

    Protocol #1: Mean Per Joint Position Error (MPJPE).
    Protocol #2: Mean Per Joint Position Error after a rigid transformation (PA MPJPE).

    ** The smaller, the better.*

  - MPI-INF-3DHP

    Percentage of Correct Keypoints (PCK). The threshold is set to 150*mm*.
    Aera under the Curve (AUC)

    ** The larger, the better.*

1. C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments,” IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 7, pp. 1325–1339, 2013.
2. D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, “Monocular 3d human pose estimation in the wild using improved cnn supervision,” in 2017 International Conference on 3D Vision (3DV). IEEE, 2017, pp. 506–516.

Experiments

Evaluation on Human3.6m

Table 1. The quantitative results compared to state-of-the-art 3d human pose estimation methods on Human3.6m.

Protocol #1	Dir.	Dis.	Eat	Gre.	Phon.	Pose	Pur.	Sit	SitD.	Smo.	Phot.	Wait	Walk	WalkD.	WalkP.	Avg
Zhou *et al.* (CVPR'16) [41]	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
Chen *et al.* (CVPR'17) [21]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1	240.1	106.7	139.2	106.2	87.0	114.1	90.6	114.2
Pavlakos *et al.* (CVPR'17) [13]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Mehta *et al.* (3DV'17) [42]	57.5	68.6	59.6	67.3	78.1	56.9	69.1	98.0	117.5	69.5	82.4	68.0	55.3	76.5	61.4	72.9
Zhou *et al.* (ICCV'17) [28]	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	51.4	63.2	55.3	64.9
Sun *et al.* (ICCV'17) [39]	52.8	54.8	54.2	54.3	61.8	67.2	53.1	53.6	71.7	86.7	61.5	53.4	61.6	47.1	53.4	59.1
Luo *et al.* (BMVC'18) [43]	53.5	60.9	56.3	59.1	64.3	74.4	55.4	63.4	74.8	98.0	61.1	58.2	70.6	49.1	55.7	63.7
Yang *et al.* (CVPR'18) [30]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Zhao *et al.* (CVPR'19) [38]	47.3	60.7	51.4	60.5	61.1	49.9	47.3	68.1	86.2	55.0	67.8	61.0	42.1	60.6	45.3	57.6
Ours	**46.0**	55.3	50.6	**53.5**	**57.5**	**46.3**	49.4	71.7	87.9	56.6	68.4	53.5	**41.4**	57.9	46.6	**56.2**

Protocol #2	Dir.	Dis.	Eat	Gre.	Phon.	Pose	Pur.	Sit	SitD.	Smo.	Phot.	Wait	Walk	WalkD.	WalkP.	Avg
Moreno-Noguer (CVPR'17) [22]	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
Sun *et al.* (ICCV'17) [39]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
Luo *et al.* (BMVC'18) [43]	40.8	44.6	42.1	45.1	48.3	54.6	41.2	42.9	55.5	69.9	46.7	42.5	48.0	36.0	41.4	46.6
Yang *et al.* (CVPR'18) [30]	26.9	30.9	36.3	39.9	43.9	47.4	28.8	29.4	36.9	58.4	41.5	30.5	29.5	42.5	32.2	**37.7**
Zhou *et al.* (TPAMI'18) [44]	47.9	48.8	52.7	55.0	56.8	65.5	49.0	45.5	60.8	81.1	53.7	51.6	54.8	50.4	55.9	55.3
Ours	35.8	41.0	42.3	42.0	**43.4**	**36.3**	36.7	55.1	66.5	**45.0**	49.6	41.2	32.9	43.9	39.0	43.4

Our method achieves
1) best performance in Protocol #1.
2) better performance than most of existing methods in Protocol #2.

Experiments

Evaluation on MPI-INF-3DHP

Table 2. The quantitative results on MPI-INF-3DHP.

Methods	Extra information	PCK	AUC
Mehta *et al.* (3DV'17) [42]		64.7	31.7
Zhou *et al.* (ICCV'17) [28]	Post-processing	68.2	32.5
Yang *et al.* (CVPR'18) [30]		69.0	32.0
Habibie *et al.* (CVPR'19) [7]	Extra training set	70.4	36.0
Want *et al.* (CVPR'19) [8]	Extra training set	81.8	54.8
Ci *et al.* (ICCV'19) [50]	2D Pose	74.0	34.7
Ours (w/o $\mathcal{L}_{OPLR}$)		23.9	8.9
Ours (full)		66.8	31.9

Table 3. The quantitative evaluation with using rigid transformation.

Methods (Using rigid transformation)	Extra information	PCK	AUC
Habibie *et al.* (CVPR'19) [7]	Extra training set	82.9	45.4
Ours		**84.4**	**46.9**

- Our method achieves superior performance even without using extra information.
- Our method outperforms the existing method with using rigid transformation for evaluation.
- Our method significantly performs better than Ours (w/o L_{OPLR}) with **an improvement from 23.9% to 66.8%**.

Experiments

Visualization

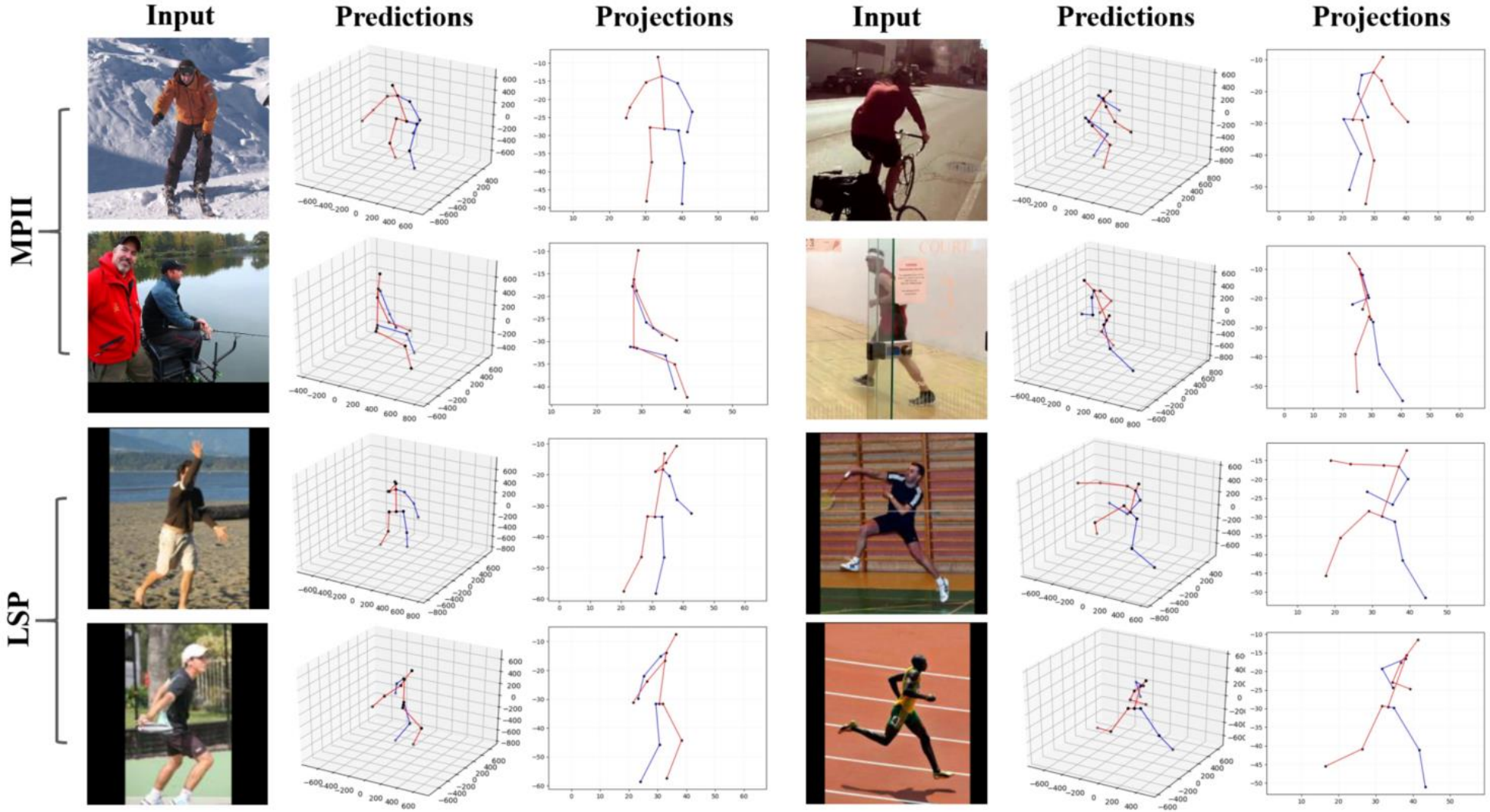

Fig. 7. The qualitative results on MPII and LSP dataset generated by the proposed method.

Summary

- We propose a novel orthographic projection and linear regression to constrain the 3D and 2D poses.

- A network is proposed which is adaptive to various in-the-wild images without retraining the 3D pose.

- Our network achieves state-of-the-art performance on the Human3.6m dataset and generalizes well to in-the-wild datasets.