Chirality Nets for Human Pose Regression

Raymond A. Yeh* Yuan-Ting Hu* Alexander G. Schwing

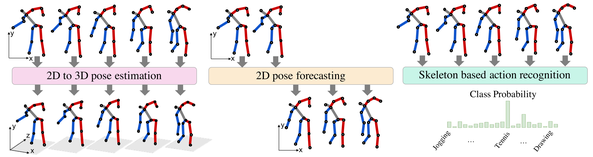
University of Illinois at Urbana-Champaign

chiralitynets.web.illinois.edu

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Human pose regression:



Prior works:

- Represent pose via a concatenation of joint coordinates
- Regression modeling with deep nets

Can we take advantage of the structure in human pose?

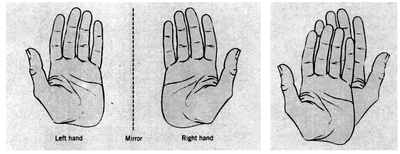
- Respect the left-right symmetry in human pose

Contributions:

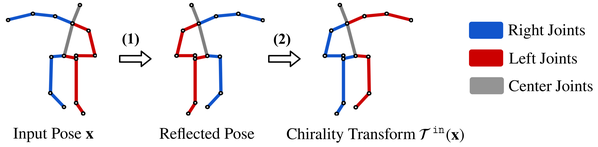
- Propose chirality nets
- Guarantee equivariance respecting left-right symmetry
- Data efficiency and reduction in computation

Introduction

Chirality: A structure and its mirror image are not superimposable, *e.g.*, left and right hand form a chiral pair

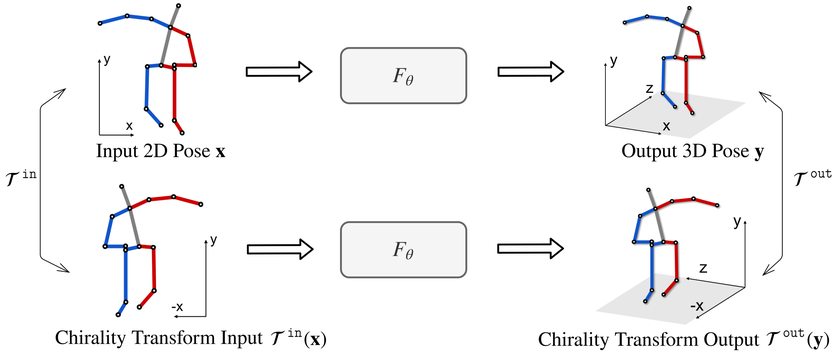


Chiral pair for human pose: Form a chiral pair via chirality transform $\mathcal{T}$



Our Approach

Equivariance to Chirality Transform:



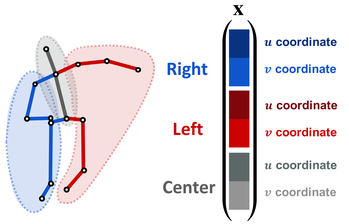
Chirality Equivariance

A function F_θ is chirality equivariant w.r.t. chiral transform $(\mathcal{T}^{\text{in}}, \mathcal{T}^{\text{out}})$ if

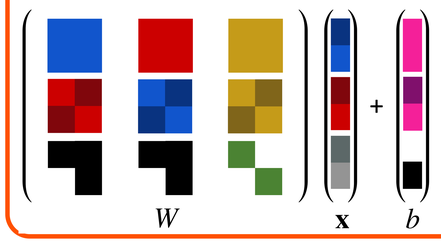
$$\mathcal{T}^{\text{out}}(F_\theta(\mathbf{x})) = F_\theta(\mathcal{T}^{\text{in}}(\mathbf{x})) \quad \forall \mathbf{x}$$

Human Pose Representation:

- Input pose $\mathbf{x} \in \mathbb{R}^{(|J_l^{\text{in}}|+|J_r^{\text{in}}|+|J_c^{\text{in}}|)\cdot|D^{\text{in}}|}$
- J_r^{in}, J_l^{in} and J_c^{in} are the set of left, right and center joints
- D^{in} denotes the dimensions per joint
  - For 2D key-points: input D^{in} consists of (u, v)
- Arrange the dimensions from left to right and u to v



Chiral fully conntected layers, $f_{\text{FC}}(\mathbf{x}; W, b) = W\mathbf{x} + b$ **:**

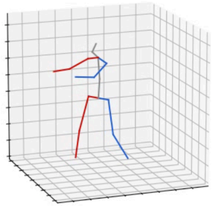


- To achieve chiral equivariance:
  (1) Odd symmetry: reflection in pose
  (2) Parameter sharing: permutation between left and right joints
- Generalize to Conv1D, LSTM, Batchnorm, *etc.*
- Reduction in parameters → better data efficiency
- Reduction in FLOPs by exploiting symmetry

Results

Video 2D to 3D Pose Estimation:

- Human3.6M Dataset
  - 17 joints as 2D keypoints
  - 17 joints as 3D keypoints
  - 15 different actions
- HumanEva-I Dataset
  - 17 joints as 2D keypoints
  - 15 joints as 3D keypoints
  - 3 different actions



Pose Estimation Results:

Results on Human3.6M (MPJPE)

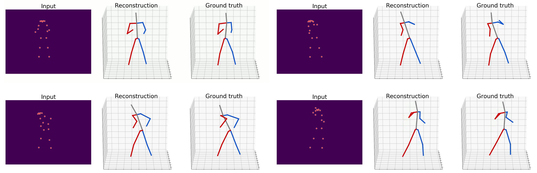
Approach	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Pavlakos (CVPR'18)	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Luvizon(CVPR'18) (⋄)	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7	61.5	70.9	53.7	48.9	57.9	44.4	48.9	53.2
Hossain (ECCV'18)(†, ⋄)	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Lee (ECCV'18)(†, ⋄)	**40.2**	49.2	47.8	52.6	50.1	75.0	50.2	**43.0**	**55.8**	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Pavllo (CVPR'19)	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
Pavllo (CVPR'19)(†)	45.9	47.5	44.3	46.4	50.0	56.9	45.6	44.6	58.8	66.8	47.9	44.7	49.7	33.1	34.0	47.7
Pavllo (CVPR'19)(†, ‡)	45.2	46.7	**43.3**	**45.6**	**48.1**	**55.1**	**44.6**	44.3	57.3	65.8	**47.1**	44.0	49.0	32.8	33.9	46.8
Ours, single-frame	47.4	49.9	47.4	51.1	53.8	61.2	48.3	45.9	60.4	67.1	52.0	48.6	54.6	40.1	43.0	51.4
Ours (†)	44.8	**46.1**	**43.3**	46.4	49.0	55.2	**44.6**	44.0	58.3	**62.7**	**47.1**	**43.9**	48.6	**32.7**	**33.3**	**46.7**

†: temporal models, ⋄: ground-truth bounding box, and ‡: test-time augmentation

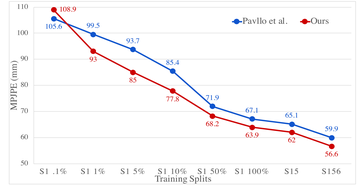
Results on Eva-I (P-MPJPE)

Approach	Walk S1	Walk S2	Walk S3	Jog S1	Jog S2	Jog S3	Box S1	Box S2	Box S3	Avg.
Pavlako(CVPR'17)	22.3	19.5	29.7	28.9	21.9	23.8	–	–	–	–
Pavlako(CVPR'18)	18.8	12.7	**29.2**	23.5	15.4	14.5	–	–	–	–
Lee(ECCV'18)	18.6	19.9	30.5	25.7	16.8	17.7	42.8	48.1	53.4	–
Pavllo(CVPR'19)	14.1	10.4	46.8	21.1	13.3	14.0	23.8	34.5	32.3	31.1
Pavllo(CVPR'19) (‡)	**13.9**	**10.2**	46.6	**20.9**	**13.1**	13.8	23.8	33.7	32.0	30.8
Ours	15.2	10.3	47.0	21.8	**13.1**	**13.7**	**22.8**	**31.8**	**31.0**	**30.6**

Qualitative Results (Boxing)



Results with limited data



Skeleton-based action recognition:

- Predicting human action from skeleton sequences
- Dataset: Kinetics-400 (Kay *et al.* 2017)
- Input: Sequence of 18 joints as 2D key-points (OpenPose)
- Output: Prediction on 400 classes

Classification Accuracy on Kinetics

Approach	Top-1	Top-5
Feature Encoding (CVPR'15)	14.9%	25.8%
Deep LSTM(CVPR'16)	16.4%	35.3%
Temporal-Conv(CVPRW'17)	20.3%	40.0%
ST-GCN (AAAI'18)	30.7%	52.8%
Ours-Conv	30.8%	52.6%
Ours-Conv-Chiral	**30.9%**	**53.0%**