Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking

- Year & Index
  - 2019
  - ArXiv:1904.01324
- Abstract excerpt
  - Pain points
    - the curse of dimen- sionality and the ill-posed nature of lifting 2D-to-3D
    - The diverse variations in background, clothing, pose, occlusions, illumination, and camera parame- ters in real-world scenarios makes it a challenging prob- lem. The popular 3D-pose annotated datasets do not cover these variations appropriately.
  - Related works
    - Lifting 2D to 3D — obtains 3D-pose from estimated 2D-pose
    - Hypothesis Generation — Some previous approaches sample multiple 3D-poses via heuristics
    - Ordinal Relations — Ordinal relations have previously been explored to estimate depth and reflectance
  - Contribution
    - · We tackle the inherent ill-posed problem of lifting 2D-to-3D human-pose by learning a deep generative model that synthesizes diverse 3D-pose samples con- ditioned on the estimated 2D-pose.
    - · We employ CVAE for 3D human-pose estimation for the first time.
    - · We derive joint-ordinal depth relations from an RGB image and employ them to rank 3D-pose samples.
    - · We show that the oracle-based pose sample obtained from our proposed generative model achieves state- of-the-art results on two benchmark datasets, Human3.6M [11] and Human-Eva [29].
    - · We show competitive performance over Baseline even when our 2D-to-3D module is trained on a separate MoCap dataset with no images.
- Method
  - 2DPoseNet: 2D-Pose from Image
  - MultiPoseNet: Multiple 3D-Poses from 2D

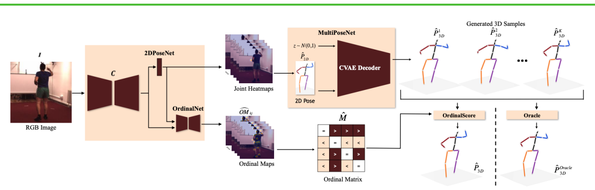

Figure 1: A schematic illustration of our modular framework. We tackle the ambiguity in the 2D-to-3D mapping by training a CVAE to generate 3D-pose samples conditioned on the 2D-pose, that are scored and weighted-averaged using joint-ordinal relations, which are regressed together with the 2D-pose. Our upper-bound performance is obtained by using an Oracle.

pipeline

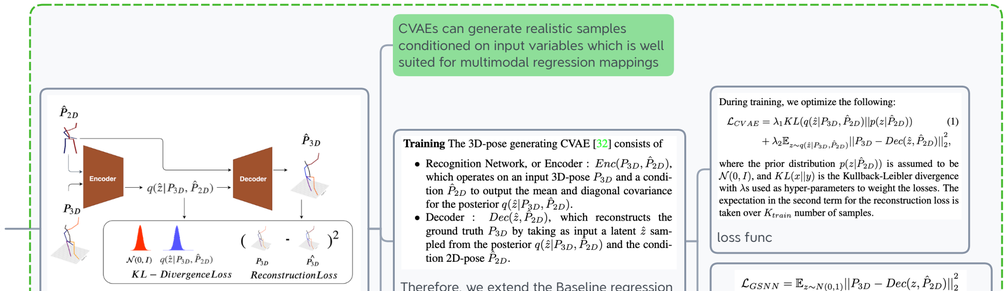

Figure 2: MultiPoseNet architecture in training. Note: in GSNN, we sample $z \sim \mathcal{N}(0, I)$ and only need the Decoder.

MultiPoseNet

CVAEs can generate realistic samples conditioned on input variables which is well suited for multimodal regression mappings

Training The 3D-pose generating CVAE [32] consists of

- Recognition Network, or Encoder : $Enc(P_{3D}, \hat{P}_{2D})$, which operates on an input 3D-pose P_{3D} and a condition $\hat{P}_{2D}$ to output the mean and diagonal covariance for the posterior $q(\hat{z}|P_{3D}, \hat{P}_{2D})$.
- Decoder : $Dec(\hat{z}, \hat{P}_{2D})$, which reconstructs the ground truth P_{3D} by taking as input a latent $\hat{z}$ sampled from the posterior $q(\hat{z}|P_{3D}, \hat{P}_{2D})$ and the condition 2D-pose $\hat{P}_{2D}$.

Therefore, we extend the Baseline regression model from into a CVAE to tackle the inherent multi-modality of the 2D-to-3D pose mapping

During training, we optimize the following:

$$\mathcal{L}_{CVAE} = \lambda_1 KL(q(\hat{z}|P_{3D}, \hat{P}_{2D})||p(z|\hat{P}_{2D})) + \lambda_2 \mathbb{E}_{z\sim q(\hat{z}|P_{3D},\hat{P}_{2D})}||P_{3D} - Dec(\hat{z}, \hat{P}_{2D})||_2^2, \quad (1)$$

where the prior distribution $p(z|\hat{P}_{2D}))$ is assumed to be $\mathcal{N}(0, I)$, and $KL(x||y)$ is the Kullback-Leibler divergence with λs used as hyper-parameters to weight the losses. The expectation in the second term for the reconstruction loss is taken over K_{train} number of samples.

loss func

$$\mathcal{L}_{GSNN} = \mathbb{E}_{z\sim N(0,1)}||P_{3D} - Dec(z, \hat{P}_{2D})||_2^2$$
$$\mathcal{L}_{hybrid} = \alpha L_{CVAE} + (1 - \alpha) L_{GSNN},$$

Combining the two we get a hybrid training objec- tive, weighted with α:

Inference We sample $z \sim \mathcal{N}(0, 1)$, and feed $(z, \hat{P}_{2D})$ to the Decoder, to obtain $\mathcal{S}_{test} = \{\hat{P}_{3D}^k; k \in \{1, 2, \ldots, K_{test}\}\}$.

inference

  - OrdinalNet: Image to Joint-Ordinal Relations

The backbone architecture for OrdinalNet is same as our 2DPoseNet

$$\hat{M}_{ij} = \begin{cases} 1 & : D_i - D_j > 0 \\ 2 & : D_i - D_j < 0 \\ 3 & : D_i - D_j \approx 0 \end{cases}$$

  - OrdinalScore: Scoring and Aggregating Generated 3D samples

$$\hat{P}_{3D} = \sum_k^{|\mathcal{S}|} p(\hat{P}_{3D}^k).\hat{P}_{3D}^k$$

pose [36]. The estimated ordinal matrix $\hat{M}$ is used to assign scores to each of the samples $P_{3D}^k \in \mathcal{S}$ by the scoring function:

$$f(\hat{P}_{3D}^k) = \sum_{i,j} \mathbb{1}(\hat{M}_{ij} == g(\hat{P}_{3D}^k)_{ij}) \quad (4)$$

$$p(\hat{P}_{3D}^k) = e^{Tf(\hat{P}_{3D}^k)} / \sum_k e^{Tf(\hat{P}_{3D}^s)}$$

  - Supervision from an Oracle — $\hat{P}_{3D}^{oracle} = \underset{s\in\mathcal{S}}{\operatorname{argmin}} \|P_{3D} - s\|_2$
- Experiment result

	Protocol 1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
PAIR	Pavlakos *et al.* [25]	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
	Zhou *et al.* [45]	54.82	60.70	58.22	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
	Martinez *et al.* [19]	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
	Sun *et al.* [34]	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
	Fang *et al.* [9]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
	*Pavlakos *et al.* [24]	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
	**Hossain *et al.*-[10]	44.2	46.7	52.3	49.3	59.9	59.4	47.5	46.2	59.9	65.6	55.8	50.4	52.3	43.5	45.1	51.9
	**Dabral *et al.*-[8]	44.8	50.4	44.7	49.0	52.9	61.4	43.5	45.5	63.1	87.3	51.7	48.5	37.6	52.2	41.9	52.1
	***Sun *et al.* [35]	47.5	47.7	49.5	50.2	51.4	43.8	46.4	58.9	65.7	49.4	55.8	47.8	38.9	49.0	43.8	49.6
	Ours (*PRED Ordinals*)	48.6	54.5	54.2	55.7	62.6	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
	Ours (*GT Ordinals*)	42.9	48.1	47.8	50.2	56.1	65.0	44.9	48.6	61.8	69.9	52.6	50.4	56.0	42.1	45.1	52.1
	Ours (*Oracle*)	37.8	43.2	43.0	44.3	51.1	57.0	39.7	43.0	56.3	64.0	48.1	45.4	50.4	37.9	39.9	46.8
UNPAIR	Martinez *et al.* [19]	109.9	112	103.8	115.3	119.3	119.3	114	116.6	118.9	127.3	112.2	119.8	113.4	119.8	111.9	115.6
	Ours (*PRED Ordinals*)	99.9	102.7	97.9	105.9	112.0	111.7	103.9	109.4	111.7	119.4	104.8	110.8	103.2	106.9	102.3	106.8
	Ours (*GT Ordinals*)	97.9	100.5	95.4	103.7	109.4	108.5	102.0	108.0	107.9	115.4	102.2	108.9	100.8	105.8	100.8	104.4
	Ours (*Oracle*)	92.6	94.6	90.6	98.4	103.8	103.3.6	96.6	101.8	101.7	108.8	96.6	102.7	95.3	100.6	96.1	98.9

Table 1: Detailed results on Human3.6M under Protocol 1(no rigid alignment in post-processing). Error is in millimeters(mm). Top: Paired methods (PAIR), Bottom: unpaired methods (UNPAIR). Results for [19] in the unpaired setting were obtained using their publicly available code. * - use additional ordinal training data from MPII and LSP. ** - use temporal information. *** - use soft-argmax for end-to-end training. These strategies are complementary with our approach.

	Protocol 2	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
PAIR	Zhou *et al.* [45]	47.9	48.8	52.7	55.0	56.8	49.0	45.5	60.8	81.1	53.7	65.5	51.6	50.4	54.8	55.9	55.3
	Pavlakos *et al.* [25]	47.5	50.5	48.3	49.3	50.7	55.2	46.1	48.0	61.1	78.1	51.1	48.3	52.9	41.5	46.4	51.9
	Martinez *et al.* [19]	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
	Fang *et al.* [9]	38.2	41.7	43.8	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
	Sun *et al.* [34]	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
	*Pavlakos *et al.* [24]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
	**Hossain *et al.*-[10]	36.9	37.9	42.8	40.3	46.8	46.7	37.7	36.5	48.9	52.6	45.6	39.6	43.5	35.2	38.5	42.0
	**Dabral *et al.*-[8]	28.0	30.7	39.1	34.4	37.1	44.8	28.9	31.2	39.3	60.6	39.3	31.1	25.3	37.8	28.4	36.3
	***Sun *et al.* [35]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.6
	Ours (*PRED Ordinals*)	35.3	35.9	45.8	42.0	40.9	52.6	36.9	35.8	43.5	51.9	44.3	38.8	45.5	29.4	34.3	40.9
	Ours (*GT Ordinals*)	31.3	31.0	39.3	37.0	37.2	47.8	32.5	32.1	39.8	47.3	40.0	34.7	41.8	27.5	31.0	36.7
	Ours (*Oracle*)	27.6	27.5	34.9	32.3	33.3	42.7	28.7	28.0	36.1	42.7	36.0	30.7	37.6	24.3	27.1	32.7
UNPAIR	Martinez *et al.* [19]	62.6	64.3	62.5	67.4	72.2	70.8	64.9	61.2	82.1	92.4	76.8	66.7	71.7	79.5	73.1	71.3
	**Ours(*PRED Ordinals*)	62.9	65.6	61.8	67.1	72.2	69.3	65.6	63.8	81.3	91.0	74.5	66.5	70.8	74.7	70.9	70.5
	Ours(*GT Ordinals*)	62.9	65.3	60.7	66.9	71.3	68.4	65.2	63.2	80.1	89.3	73.5	66.1	70.5	74.7	70.9	70.0
	Ours (*Oracle*)	56.8	59.2	55.0	59.6	65.6	62.0	58.4	56.5	74.2	82.8	67.6	60.0	63.6	68.2	64.3	63.6

Table 2: Detailed results on Human3.6M under Protocol 2(rigid alignment in post-processing). Top: Paired methods (PAIR), Bottom: unpaired methods (UNPAIR). Results for [19] in the unpaired setting were obtained using their publicly available code.