Ordinal Depth Supervision for 3D Human Pose Estimation

Year & Index

- 2018
- ArXiv:1805.04095

Abstract excerpt

Pain points

- the need for accurate 3D ground truth
- limited availability of 3D annotations for natural im- ages.
  - Most datasets are captured using Motion Capture (MoCap) systems in a studio setting and it is difficult to reach the variability of 2D human pose datasets, like MPII or LSP.

Related works

Reconstruction approaches

follows the reconstruction paradigm by employing 2D pose detectors to localize 2D human joints and using these lo- cations to estimate plausible 3D poses

- Despite the success of this paradigm, it comes with important drawbacks.
  - No image- based evidence is used during the reconstruction step,
  - the result is too reliant on an imperfect 2D pose detector and even for perfect 2D correspondences,
  - the 3D estimate might fail because of the reconstruction ambiguity.
- In contrast, by using ordinal depth relations we can leverage rich image- based information during estimation, without relinquishing the accuracy of reconstruction approaches, which can also be integrated in our framework (Section 3.4).

Discriminative approaches

Discriminative approaches are orthogonal to the reconstruction paradigm since they estimate the 3D pose directly from the image.

- The main critique of these end-to-end approaches is that images with corresponding 3D ground truth are required for train- ing.
- Our work attempts to relax this important constraint, by training with weak 3D information in the form of ordi- nal depth relations for the joints and 2D keypoints

Generating training examples

- The limited availability of 3D ground truth for training 3D human pose ConvNets
- use graphics to augment the training data
- composing hu- man parts from different images to produce combinations with known 3D pose
- large-scale data collection is not trivial

3D annotations

Prior works have also relied on humans to perceive and annotate 3D properties that are lost through the projection of a 3D scene on a 2D image.

- laborious procedure

we

- provide annotations by humans for a much larger set of images (i.e., more than 15k images with our annotations compared to 1k for the posebits dataset),
- and instead of exploring an extensive set of pose attributes, we propose a cleaner training scheme that requires only 2D keypoint locations and ordinal depth relations.

Ordinal relations

- many algorithms for learning-to-rank have been developed
- In the context of computer vi- sion, previous works have used relations to learn apparent depth [69, 12] or reflectance [29, 63] of a scene. We share a common motivation with these appr
- We share a common motivation with these approaches in the sense that ordinal relations are easier for humans to annotate, com- pared to metric depth or absolute reflectance values.

Contribution

- We propose the use of ordinal depth relations of human joints for 3D human pose estimation to bypass the need for accurate 3D ground truth.
  - ordinal annotations can be easily incorporated in the training procedure of typical ConvNets for 3D human pose.
- We showcase the flexibility of the ordinal relations by incorporating them in different network settings, where we always achieve competitive results to train- ing with the actual 3D ground truth.
- We augment two popular 2D pose datasets (LSP and MPII) with ordinal depth annotations and demonstrate the applicability of the proposed approach to 3D pose estimation in non-studio conditions.
- We include our ordinal annotations in the training pro- cedure of typical ConvNets for 3D human pose and exemplify their effectiveness by achieving new state- of-the-art results on the standard benchmarks.

Design

Depth prediction

instead of explicitly predicting the 3D pose, we only predict depth values for the joints.

Let us represent the human body with N joints. For each joint i we want to predict its depth zi. The provided data are in the form of pairwise ordinal depth relations. For a pair of joints (i, j), we denote the ordinal depth relation as r(i,j) taking the value:

r(i,j)

- · +1, if joint i is closer than j,
- · −1, if joint j is closer than i,
- · 0, if their depths are roughly the same.

The ConvNet we use for this task takes the image as input and predicts N depth values zi, one for each joint. Given the r(i,j) relation and assuming that the ConvNet is produc- ing the depth estimates zi and zj for the two corresponding joints, the loss for this pair is:

differentiable ranking loss expression,

Li,j =

- log(1+exp(zi −zj)), r(i,j) =+1
- log(1+exp(−zi +zj)), r(i,j) =−1
- (zi −zj)2, r(i,j) =0.

Lrank = sum Li,j. (i,j)∈I

- The loss can be computed based only on the subset of pairs that have been annotated.
- Additionally, the relations do not have to be consistent, i.e., no strict global ordering is required.
- Instead, the ConvNet is allowed to learn a consensus from the provided relationships by minimizing the incurred loss.

Intuitively, it enforces a large margin between the values zi and zj if one of them has been annotated as closer than the other, otherwise it enforces them to be equal.

Coordinate prediction for 3D pose

Given the ConvNet used in the previous sec- tion, the most natural extension is to enrich its output by predicting the 2D coordinates of the joints as well. Thus, we predict 2N additional values which correspond to the pixel coordinates w = (x, y) of each joint.

weakly supervised version.

$$\mathcal{L}_{rank} = \sum_{(i,j)\in\mathcal{I}} \mathcal{L}_{i,j}.$$

L_{rank}:ranking loss for the values zn

$$\mathcal{L}_{i,j} = \begin{cases} \log(1+\exp(z_i - z_j)), & r_{(i,j)} = +1 \\ \log(1+\exp(-z_i + z_j)), & r_{(i,j)} = -1 \\ (z_i - z_j)^2, & r_{(i,j)} = 0. \end{cases}$$

loss_ij

$$\mathcal{L}_{keyp} = \sum_{n=1}^{N} \|\boldsymbol{w}_n - \hat{\boldsymbol{w}}_n\|_2^2.$$

the regression loss for the keypoint coordinates wn

$$\mathcal{L} = \mathcal{L}_{rank} + \lambda \mathcal{L}_{keyp}$$

where the value λ = 100 is used for our experiments.

Volumetric prediction for 3D pose

Precisely, for every joint n, the ConvNet predicts score maps Ψn, which can be transformed to a probability distribution, by apply- ing a softmax operation σ. So, the joint n is located in posi- tion u = (x, y, z) with probability p(u|n) = σ[Ψn]u. The marginalized probability distribution in the 2D plane is:

$$p(x, y|n) = \sum_z p(\boldsymbol{u}|n),$$

Similarly, the marginalized probability distribution for the depth dimension is:

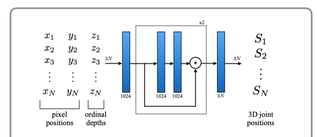
$$p(z|n) = \sum_{x,y} p(\boldsymbol{u}|n),$$

Our loss function takes the form:

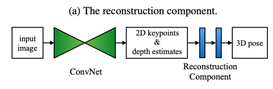
L = Lrank + λLheat

- The loss for the z-dimension, Lrank, is the same ranking loss as before
  - where we recover depth for each joint by taking the mean value of the estimated soft distribution: zn = sum_z zp(z|n).
- Lheat is an L2 loss between the predicted and the ground truth heatmaps
  - For the x-y dimensions, the target for each keypoint is a heatmap with a Gaussian centered around its ground truth location and

Integration with a reconstruction component



(a) The reconstruction component.



(b) Integration of the reconstruction component.

Figure 3: (a) The reconstruction component is a multi-layer perceptron with two bilinear units [25]. The input is the concatenation of the pixel locations of the joints (x_i, y_i), and the ordinal depths z_i, while the output is the 3D pose coordinates S_i. (b) Integration of the reconstruction module in the full framework. The ConvNet of Section 3.2 or 3.3 estimates 2D keypoint locations and depths which are used by the reconstruction module to predict a coherent 3D pose.

denoting with Si hat the output 3D joints of the ConvNet and with Si the joints of the 3D pose that was used to generate the input,

$$\mathcal{L}_{3D} = \sum_{n=1}^{N} \|S_n - \hat{S}_n\|_2^2.$$

Experiment result

Architecture		Supervision	Avg error
depth prediction		ordinal supervision	84.24
		direct regression	80.23
coordinate regression		weakly supervised	115.08
		fully supervised [32]	112.41
volume regression	one hourglass	weakly supervised	89.93
		fully supervised [32]	85.82
	two hourglasses	weakly supervised	79.03
		fully supervised [32]	69.77

Table 1: Effect of training with the actual 3D ground truth, versus employing weaker ordinal depth supervision on Hu- man3.6M. The results are mean per joint errors (mm).

Effect of training with the actual 3D ground truth, versus employing weaker ordinal depth supervision on Hu- man3.6M

	Avg error
Human3.6M	71.9
Human3.6M + 2D keyp	66.6
Human3.6M + 2D keyp + Ord	62.1
Human3.6M + 2D keyp + Rec	59.1
Human3.6M + 2D keyp + Ord + Rec	**56.2**

Table 2: Ablative study on Human3.6M demonstrating the effect of incorporating additional data sources in the train- ing procedure (2D keypoints and ordinal depth relations), as well as integrating a rconstruction component. The num- bers are mean per joint errors (mm).

Ablative study on Human3.6M demonstrating the effect of incorporating additional data sources in the train- ing procedure (2D keypoints and ordinal depth relations), as well as integrating a rconstruction component.

	PCK3D	AUC
Human3.6M	17.1	6.3
Human3.6M + 2D keyp	44.3	19.8
Human3.6M + 2D keyp + Ord	**71.9**	**35.3**

Table 3: Ablative study on MPI-INF-3DHP demonstrating that supervision through our ordinal annotations is impor- tant for proper generalization.

Ablative study on MPI-INF-3DHP demonstrating that supervision through our ordinal annotations is impor- tant for proper generalization.

	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Tekin *et al.* [49] (CVPR'16)	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.* [67] (CVPR'16)	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.* [41] (ECCV'16)	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
Zhou *et al.* [65] (ECCVW'16)	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
Chen *et al.* [10] (CVPR'17)	89.9	97.6	90.0	107.9	107.3	139.2	93.6	136.1	133.1	240.1	106.7	106.2	114.1	87.0	90.6	114.2
Tome *et al.* [51] (CVPR'17)	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8	110.2	173.9	85.0	85.8	86.3	71.4	73.1	88.4
Rogez *et al.* [40] (CVPR'17)	76.2	80.2	75.8	83.3	92.2	105.7	79.0	71.7	105.9	127.1	88.0	83.7	86.6	64.9	84.0	87.7
Pavlakos *et al.* [32] (CVPR'17)	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
Nie *et al.* [60] (ICCV'17)	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
Tekin *et al.* [48] (ICCV'17)	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	74.3	69.7
Zhou *et al.* [64] (ICCV'17)	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.2	66.1	51.4	63.2	55.3	64.9
Martinez *et al.* [25] (ICCV'17)	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
Ours	**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**

Table 4: Detailed results on Human3.6M [16]. Numbers are mean per joint errors (mm). The results of all approaches are obtained from the original papers. We outperform all other approaches across the table.

	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Akhter & Black [1]* (CVPR'15)	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.* [38]* (ECCV'12)	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.* [66]* (CVPR'15)	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.* [5] (ECCV'16)	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 [28] (CVPR'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
Pavlakos *et al.* [32] (CVPR'17)	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.* [25] (ICCV'17)	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
Ours	**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**

Table 5: Detailed results on Human3.6M [16]. Numbers are reconstruction errors. The results of all approaches are obtained from the original papers, except for (*), which were obtained from [5]. We outperform all other approaches across the table.

Human3.6M:
mean per joint errors (mm)
reconstruction errors

	Walking			Jogging			
	S1	S3	S3	S1	S2	S3	Avg
Radwan *et al.* [37]	75.1	99.8	93.8	79.2	89.8	99.4	89.5
Wang *et al.* [54]	71.9	75.7	85.3	62.6	77.7	54.4	71.3
Simo-Serra *et al.* [43]	65.1	48.6	73.5	74.2	46.6	32.2	56.7
Bo *et al.* [4]	46.4	30.3	64.9	64.5	48.0	38.2	48.7
Kostrikov *et al.* [20]	44.0	30.9	41.7	57.2	35.0	33.3	40.3
Yasin *et al.* [62]	35.8	32.4	41.6	46.6	41.4	35.4	38.9
Moreno-Noguer [28]	19.7	13.0	**24.9**	39.7	20.0	21.0	26.9
Pavlakos *et al.* [32]	22.1	21.9	29.0	29.8	23.6	26.0	25.5
Martinez *et al.* [25]	19.7	17.4	46.8	26.9	18.2	18.6	24.6
Ours	**18.8**	**12.7**	29.2	**23.5**	**15.4**	**14.5**	**18.3**

Table 6: Results on the HumanEva-I [42] dataset. Num- bers are reconstruction errors (mm). The results of all ap- proaches are obtained from the original papers.

HumanEva-I

Approach	Studio GS	Studio no GS	Outdoor	All	All
	3DPCK	3DPCK	3DPCK	3DPCK	AUC
Mehta *et al.* [26]	70.8	62.3	58.8	64.7	31.7
Zhou *et al.* [64]	71.1	**64.7**	72.7	69.2	32.5
Ours	**76.5**	63.1	**77.5**	**71.9**	**35.3**

Table 7: Detailed results on the test set of MPI-INF-3DHP [26]. The results for all approaches are taken from the original papers. No training data from this dataset have been used for training by any method.

MPI-INF- 3DHP