Recurrent 3D Pose Sequence Machines

Mude Lin, Liang Lin, Xiaodan Liang, Keze Wang, Hui Cheng

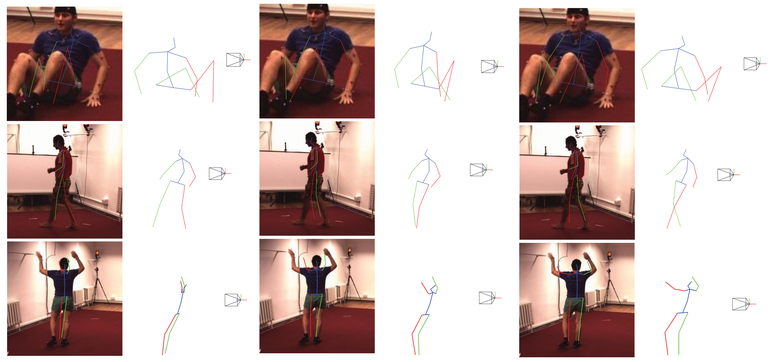
School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 2017 HONOLULU

Introduction

- Challenges:
  - Large variation in human appearance,
  - Arbitrary camera viewpoints and obstructed visibilities due to external entities and self-occlusions.
  - 3D pose is inherently ambiguous from a geometric perspective
- Main contributaions:
  - We propose a novel RPSM model that learns to recurrently integrate rich spatial and temporal long-range dependencies using a multi-stage sequential refinement, instead of relying on manually defined body smoothness or kinematic constraints. (Fig. 1)
  - Casting the recurrent network models to sequentially incorporate 3D pose geometry structural information is innovative in literature, which may also inspire other 3D vision tasks.
  - Extensive evaluations on the public challenging Human3.6M dataset and HumanEva-I dataset show that our approach outperforms existing methods of 3D human pose estimation by large margins.



(a) 1st-stage (b) 2nd-stage (3) 3rd-stage

Figure 1: Some visual results of our approach (RPSM) on Human3.6M dataset. The estimated 3D skeletons are reprojected into the images and shown by themselves from the side view (next to the images). The figures from left to right correspond to the estimated 3D poses generated by the 1st-stage, 2nd-stage and 3rd-stage of RPSM, respectively.

RPSM Architecture

- We propose a novel Recurrent 3D Pose Sequence Machine (RPSM) for estimating 3D human poses from a sequence of images. Inspired by convolutional pose machine [34] architectures for 2D pose estimation, our RPSM proposes a multi-stage training to capture long-range dependencies among multiple body-parts for 3D pose prediction, and further enforce the temporal consistency between the predictions of sequential frames. (Fig. 2)
- At each stage, our RPSM is composed by a 2D pose module, a feature adaption module, and a 3D pose recurrent module. (Fig. 3)

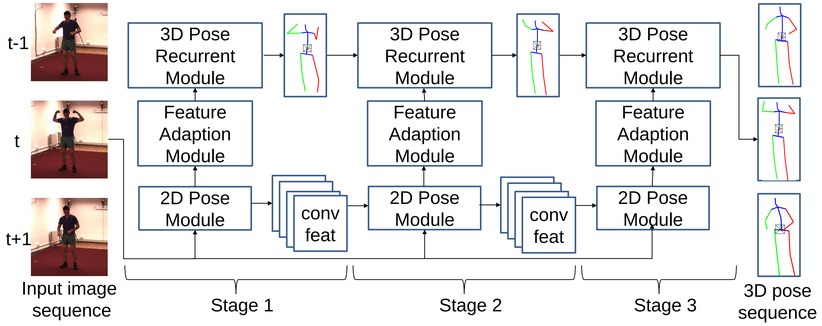


Figure 2: An overview of the proposed RPSM architecture.

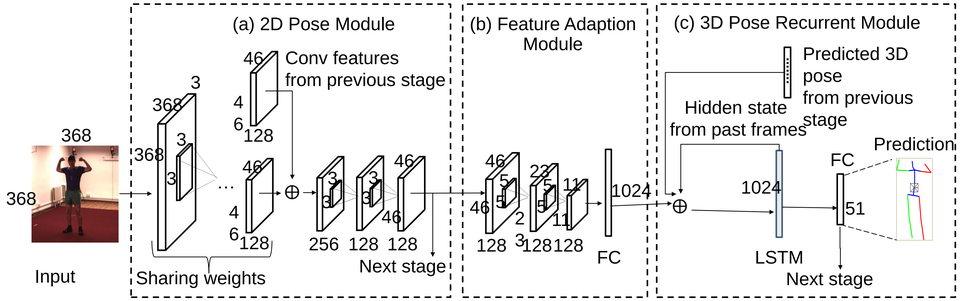


Figure 3: Detailed network architecture of our proposed RPSM at the k-th stage.

Results

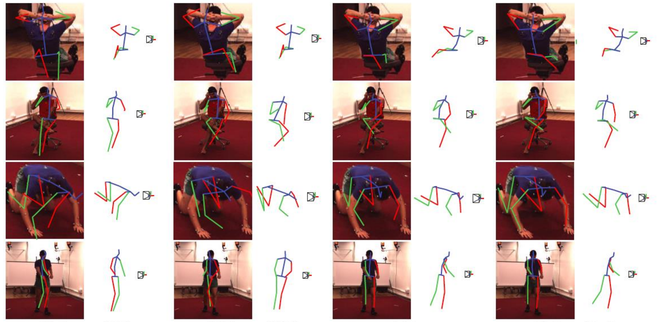
- We perform the extensive evaluations on two publicly available datasets: Human3.6M [16] and HumanEva-I [25]

Method	Direction	Discuss	Eating	Greet	Phone	Pose	Purchase	Sitting	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg.
LinKDE [16]	132.71	183.55	132.37	164.39	162.12	150.61	171.31	151.57	243.03	162.14	205.94	170.69	96.60	177.13	127.88	162.14
Li *et al.* [20]	-	136.88	96.94	124.74	-	-	-	-	-	-	168.68	-	69.97	132.17	-	-
Tekin *et al.* [30]	102.39	158.52	87.95	126.83	118.37	114.69	107.61	136.15	205.65	118.21	185.02	146.66	65.86	128.11	77.21	125.28
Zhou *et al.* [41]	87.36	109.31	87.05	103.16	116.18	106.88	99.78	124.52	199.23	107.42	143.32	118.09	79.39	114.23	97.70	113.01
Zhou *et al.* [40]	91.83	102.41	96.95	98.75	113.35	90.04	93.84	132.16	158.97	106.91	125.22	94.41	79.02	126.04	98.96	107.26
Du *et al.* [9]	85.07	112.68	104.90	122.05	139.08	105.93	166.16	117.49	226.94	120.02	135.91	117.65	99.26	137.36	106.54	126.47
Sanzari *et al.* [24]	**48.82**	**56.31**	95.98	84.78	96.47	66.30	107.41	116.89	129.63	97.84	105.58	**65.94**	92.58	130.46	102.21	93.15
Ours	58.02	68.16	**63.25**	**65.77**	**75.26**	**61.16**	**65.71**	**98.65**	**127.68**	**70.37**	**93.05**	68.17	**50.63**	**72.94**	**57.74**	**73.10**

Table 1: Quantitative comparisons on Human3.6M dataset using 3D pose errors (in millimeter) for different actions of subjects 9 and 11.

Methods	Walking				Jogging				Boxing			
	S1	S2	S3	Avg.	S1	S2	S3	Avg.	S1	S2	S3	Avg.
Simo-Serra *et al.* [28]	99.6	108.3	127.4	111.8	109.2	93.1	115.8	108.9	-	-	-	-
Radwan *et al.* [21]	75.1	99.8	93.8	89.6	79.2	89.8	99.4	89.5	-	-	-	-
Wang *et al.* [31]	71.9	75.7	85.3	77.6	62.6	77.7	54.4	71.3	-	-	-	-
Du *et al.* [9]	62.2	61.9	69.2	64.4	56.3	59.3	59.3	58.3	-	-	-	-
Simo-Serra *et al.* [27]	65.1	48.6	73.5	62.4	74.2	46.6	32.2	56.7	-	-	-	-
Bo *et al.* [3]	45.4	28.3	62.3	45.3	55.1	43.2	37.4	45.2	42.5	64.0	69.3	58.6
Kostrikov *et al.* [18]	44.0	30.9	41.7	38.9	57.2	35.0	33.3	40.3	-	-	-	-
Tekin *et al.* [29]	37.5	25.1	49.2	37.3	-	-	-	-	50.5	61.7	**57.5**	56.6
Yasin *et al.* [36]	35.8	32.4	41.6	36.6	46.6	41.4	35.4	38.9	-	-	-	-
Ours	**26.5**	**20.7**	**38.0**	**28.4**	**41.0**	**29.7**	**29.1**	**33.2**	**39.4**	**57.8**	61.2	**52.8**

Table 2: Quantitative comparisons on HumanEva-I dataset using 3D pose errors (in millimeter) for the “Walking”, “Jogging” and “Boxing” sequences.



(a) Zhou *et al.* [39] (b) Zhou *et al.* [38] (c) Ours (d) Ground Truth

Figure 4: Empirical study on the qualitative comparisons on Human3.6M dataset. The 3D pose are visualized from the side view and the camera are also depicted.

Code available at