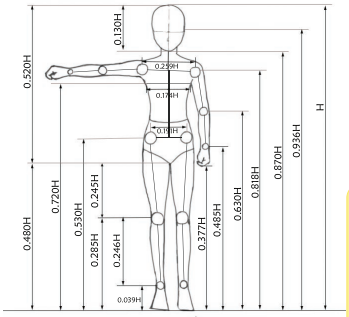
Marker-less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps

Yu Du[1], Yongkang Wong[2], Yonghao Liu[1], Feilin Han[1],Yilin Gui[1], Zhen Wang[1], Mohan Kankanhalli[2], Weidong Geng[1⋆]

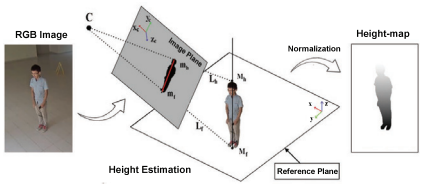
[1]Zhejiang University, [2]National University of Singapore

Motivation



The recovery of 3D human pose with monocular camera is an inherently ill-posed problem due to the large number of possible projections from the same 2D image to 3D space.

- Human observers are able to accurately estimate the pose of a human body with a single eye by leveraging vast memories of **anatomical structure of human body**.
- The difference of the recovered human poses in consecutive frames should **be consistent with the actual velocity of each joint**.



Contributions

- The RGB image and its calculated **height-map** are combined to detect the landmarks of 2D joints with a dual-stream ConvNet.
- We formulate a new objective function to estimate 3D motion from the detected 2D joints in the monocular image sequence, which reinforces the temporal coherence constraints on both the camera and 3D poses.

Framework

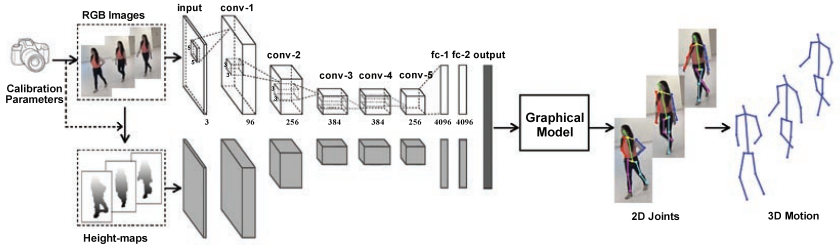


Image → 2D: minimize the energy over the locations $\boldsymbol{l}$ and types $\boldsymbol{t}$ of joints

$$F(\boldsymbol{l},\boldsymbol{t}|\boldsymbol{I}) = \underbrace{\sum_{i\in\mathcal{V}} U(\boldsymbol{l}_i|\boldsymbol{I})}_{\text{The energy of a joint locating at certain position}} + \underbrace{\sum_{(i,j)\in\mathcal{E}} R(\boldsymbol{l}_i,\boldsymbol{l}_j,t_{ij},t_{ji}|\boldsymbol{I})}_{\text{The energy of two neighboring joints has certain relationship}} + w_0$$

2D → 3D: minimize the loss over the coefficients of a dictionary of 3D pose $\mathbf{P}$ and its **pose-conditioned joint velocity V**.

$$\min_{\theta} \overbrace{\mathcal{L}(\theta;\mathbf{p})}^{\text{3D to 2D } (\mathbf{P}\rightarrow\mathbf{p}) \text{ projection error}} + \underbrace{\mathcal{R}_t(\theta)}_{\text{Continuity constraints of poses } (\mathbf{P}) \text{ and camera } (\mathbf{R} \text{ and } \mathbf{T})} + \overbrace{\mathcal{R}_a(\theta)}^{\text{Length constraints of arms and legs}}$$

$\mathcal{R}_t(\theta) = \alpha\|\nabla_t(\mathbf{P}-\mathbf{V})\|^2 + \beta_r\|\nabla_t\mathbf{R}\|^2 + \beta_t\|\nabla_t\mathbf{T}\|^2$

Training of the dual-stream ConvNet: The RGB stream is pre-trained on LSP dataset, and the resultant network is further applied on our synthetic height-maps dataset to obtain the initial weights of the height stream. The entire network is then jointly fine-tuned on a target training set.

Evaluation of 2D Joints Localization

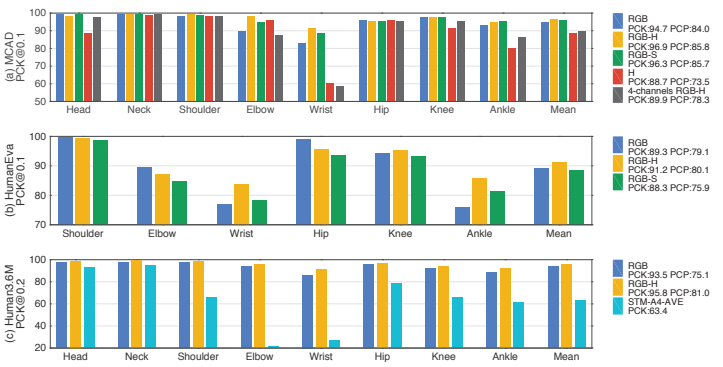


Figure 1: Evaluation of 2D joints localization with RGB [2], **RGB-H (RGB-Height maps)**, RGB-S (RGB-Silhouette), H (height-maps), 4-channels RGB-H and STM-A4-AVE [44] respectively on the MCAD, HumanEva and Human3.6M.

Evaluation of 3D Motion Recovery with Ground-Truth 2D Joints

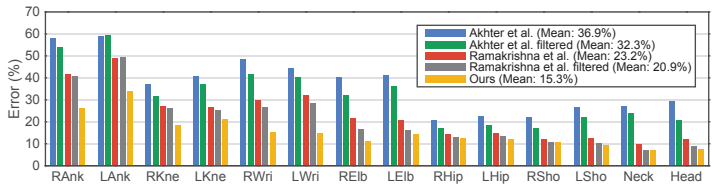
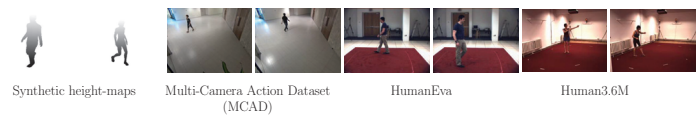


Figure 2: Evaluation of 3D motion recovery with known 2D joints. The respective average error is shown in the legend. The estimated poses of [23] and [9] are further filtered by zero-phase Butterworth filter (3rd order, 0.2 Hz for [23]; 2nd order, 1.7 Hz for [9]).

Datasets



Evaluation of 3D Motion Recovery with Predicted 2D Joints

	Walking				Jogging			
	S1	S2	S3	**Mean**	S1	S2	S3	**Mean**
[10]	99.6 (42.6)	108.3 (42.3)	127.4 (24.0)	111.8	109.2 (41.5)	93.1 (41.1)	115.8 (40.6)	106.0
[5]	71.9 (19.0)	75.7 (15.9)	85.3 (10.3)	77.6	62.6 (10.2)	77.7 (12.1)	**54.4** (9.0)	64.9
Ours	**62.2** (18.6)	**61.9** (13.2)	**69.2** (22.4)	**64.4**	**56.3** (15.4)	**59.3** (14.4)	59.3 (15.5)	**58.3**

Table 1: Evaluation of 3D motion estimation on 3 subjects of the HumanEva dataset. The value in each cell are the RMS error and standard deviation in millimeter.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases	Sitting
LinKDE [45]	132.71	183.55	132.37	164.39	162.12	205.94	150.61	171.31	151.57
Li *et al.* [28]	-	136.88	**96.94**	124.74	-	168.68	-	-	-
Ours	**85.07**	**112.68**	104.90	**122.05**	**139.08**	**135.91**	**105.93**	**166.16**	**117.49**

	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	**Avg (6 actions)**	**Avg (15 actions)**
LinKDE [45]	243.03	162.14	170.69	177.13	96.60	127.88	160.00	162.14
Li *et al.* [28]	-	-	-	**132.17**	**69.97**	-	121.56	-
Ours	**226.94**	**120.02**	**117.65**	137.36	99.26	**106.54**	**118.69**	**126.47**

Table 2: Evaluation of 3D motion estimation on Human3.6M dataset. The error are reported in mean per joint position error (MPJPE) [45].

References

[2] Chen, X., Yuille, A.L.: Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. NIPS (2014)
[5] Wang, C., *et al.*: Robust Estimation of 3D Human Poses from a Single Image. CVPR (2014)
[9] Ramakrishna, V., *et al.*: Reconstructing 3D Human Pose from 2D Image Landmarks. ECCV (2012)
[10] Simo-Serra, E., *et al.*: Single image 3D human pose estimation from noisy observations. CVPR (2012)
[23] Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. CVPR (2015)
[28] Li, S., *et al.*: Maximum-margin structured learning with deep networks for 3D human pose estimation. ICCV (2015)
[44] Zhou, F., la Torre, F.D.: Spatio-temporal matching for human pose estimation in video. PAMI (2016)
[45] Ionescu, C., *et al.*: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. PAMI (2014)

ZHEJIANG UNIVERSITY

NUS National University of Singapore

Scan for more details