Fusing Wearable IMUs with Multi-View Images for Human Pose Estimation: A Geometric Approach

Zhe Zhang[1,2], Chunyu Wang[2], Wenhu Qin[1], Wenjun Zeng[2]

[1]Southeast University, [2]Microsoft Research Asia

Introduction

The Task

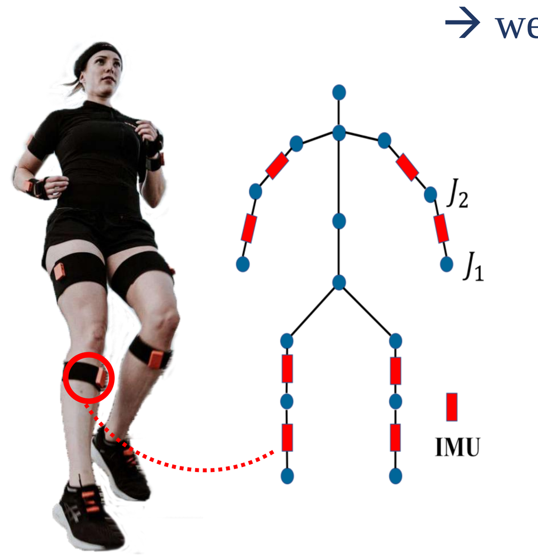
recovering absolute 3D human pose in world coordinate system by fusing *Wearable IMUs* and *Multi-View Images*

Previous Methods

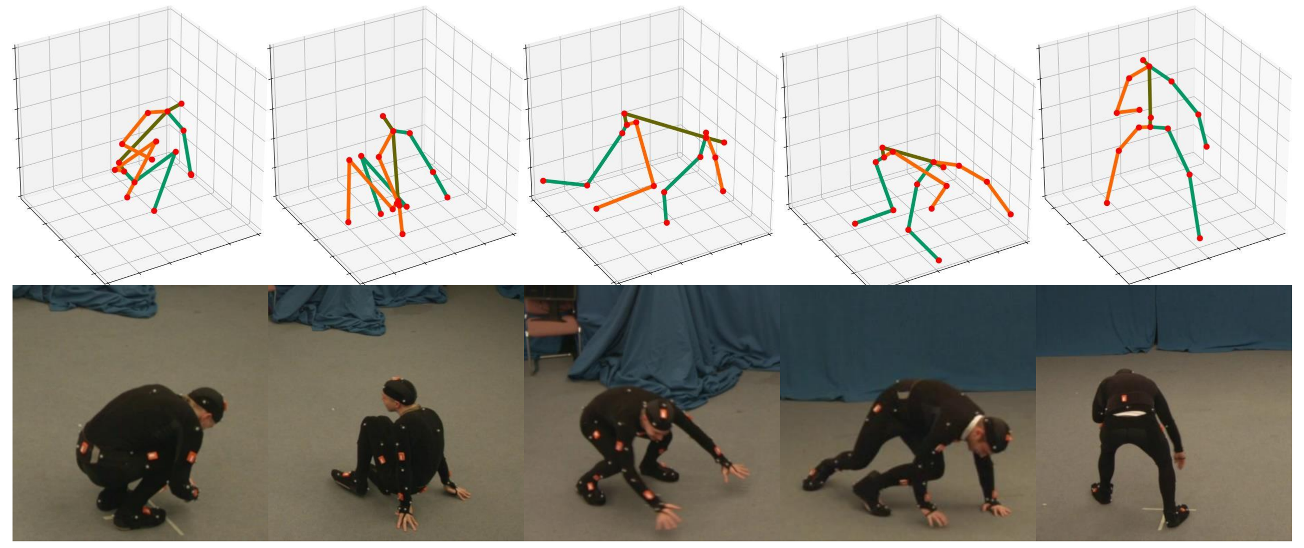
- Optimization based: estimate 3D human pose by *minimizing an energy function* which is related to both IMUs and image features
- Ad-Hoc method: estimate 3D poses *separately* from the images and IMUs, and then combine them to get the final estimation

Main Challenges

It is nontrivial to deeply and effectively incorporate IMUs in the existing image processing pipeline



IMG © Xsens



Contribution

Cross-Joint-Fusion in both **2D & 3D** pose estimation

❖ **Orientation Regularized Network** (ORN)

- IMU orientations as a structural prior
- mutually fuse the image features of each pair of joints linked by IMUs
- For example, it uses the features of the elbow to reinforce those of the wrist based on the IMU at the lower-arm.

❖ **Orientation Regularized Pictorial Structure Model** (ORPSM)

- an orientation prior that requires the limb orientations of the 3D pose to be consistent with the IMUs

SOTA Results

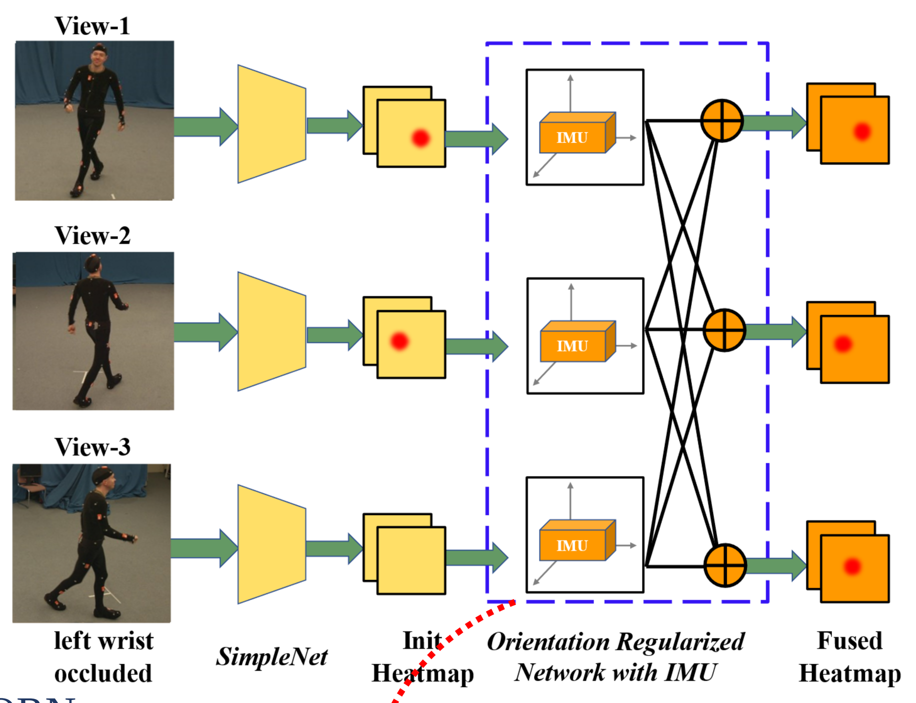
- final 3D pose error is significantly smaller than previous SOTAs on Total Capture Datasets
- proof-of-concept analysis on Human3.6M Dataset by synthesizing IMUs from ground-truth

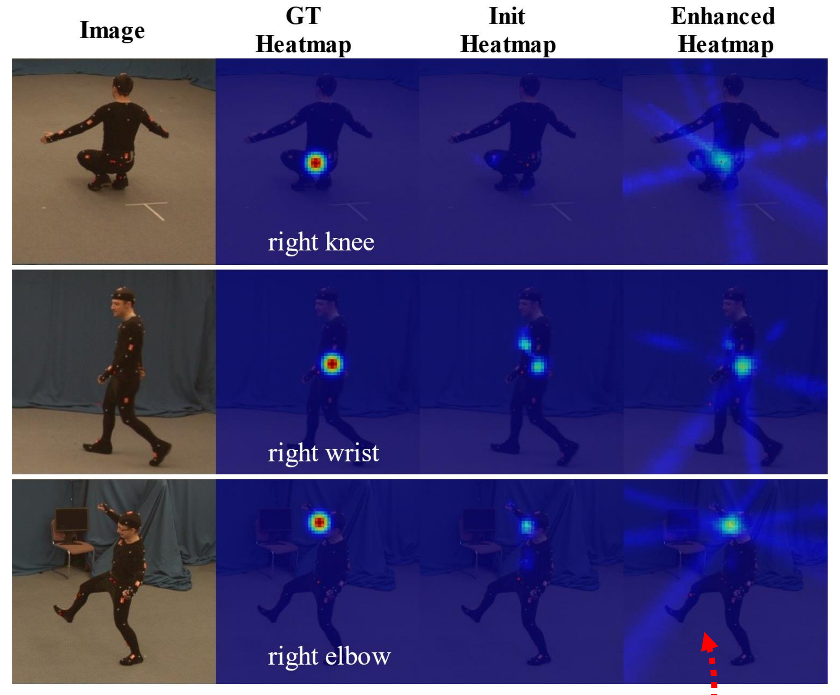
CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Orientation Regularized Network (ORN)

determining the *relative positions* between each *pair of joints* in the images is challenging

→ we solve elegantly in the 3D space with the help of IMU orientations





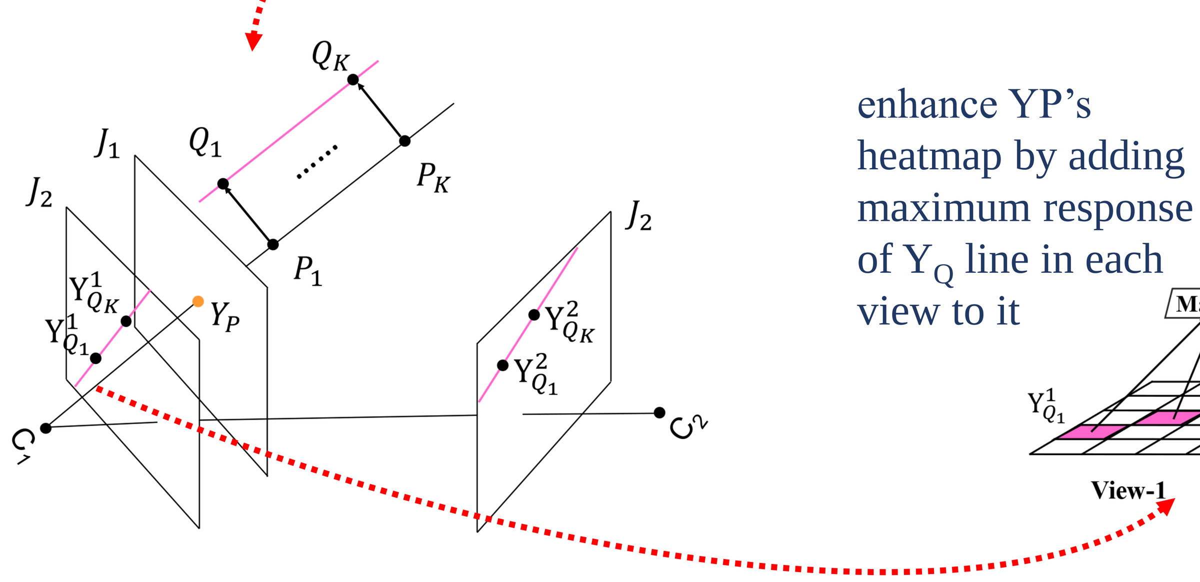
correct location will be enhanced most

Main Challenge in ORN

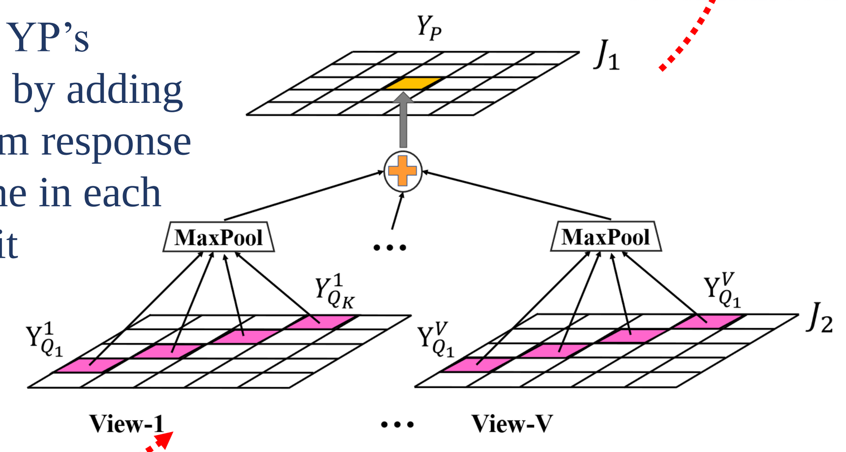
- depth is an ambiguity
- determine relative positions between each pair of joints (Y_P and Y_Q) in the images

Solution

find all possible Y_Q corresponding to Y_P in a line by adding limb offset (*orient * length*)



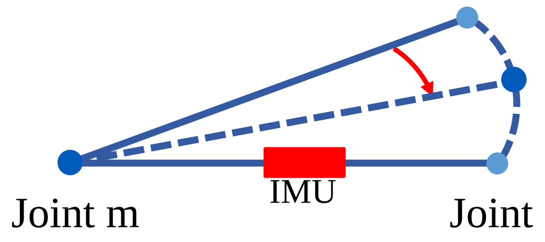
enhance YP's heatmap by adding maximum response of Y_Q line in each view to it

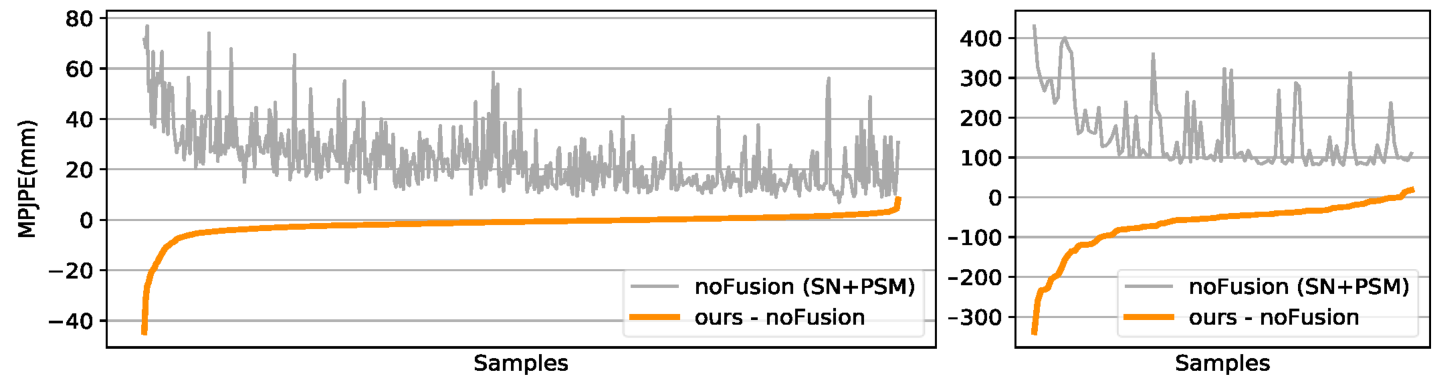


Orientation Regularized PSM (ORPSM)

- pictorial model is used to estimate 3D pose
- dot product between the *limb orientations of the estimated pose* and the *IMU orientations* as the limb orientation potential
- works as a soft constraint to let limb comply to IMU orientations

$$\psi^{IMU}(J_m, J_n) = \frac{J_m - J_n}{|J_m - J_n|_2} \cdot o_{m,n}$$





grey line:
3D MPJPE error of *noFusion* approach

orange line:
error difference between our method and noFusion

orange line below zero
→ *our method* has smaller errors

Code released at: *aka.ms/imu-human-pose*

Experimental Results

Table 1. The 2D pose estimation accuracy (PCKh@t) on the Total Capture Dataset

Methods	PCKh @	Hip	Knee	Ankle	Shldr	Elbow	Wrist	Mean (Six)	Others	Mean (All)
SN	1/2	99.3	98.3	98.5	98.4	96.2	95.3	97.7	99.5	98.1
ORN^{same}	1/2	99.4	99.0	98.8	98.5	97.7	96.7	98.3	99.5	98.6
ORN	1/2	**99.6**	**99.2**	**99.0**	**98.9**	**98.0**	**97.4**	**98.7**	**99.5**	**98.9**
SN	1/6	97.5	92.3	92.5	78.3	80.8	80.0	86.9	95.4	89.1
ORN^{same}	1/6	97.2	94.0	93.3	78.1	83.5	82.0	88.0	95.4	89.9
ORN	1/6	**97.7**	**94.8**	**94.2**	**81.1**	**84.7**	**83.6**	**89.3**	**95.4**	**90.9**
SN	1/12	87.6	67.0	68.6	47.4	50.0	49.3	61.7	78.1	65.8
ORN^{same}	1/12	81.2	70.1	68.0	43.9	51.6	50.1	60.8	78.1	65.2
ORN	1/12	**85.3**	**71.6**	**70.6**	**47.7**	**53.2**	**51.9**	**63.4**	**78.1**	**67.1**

Table 2. 3D pose estimation errors (mm) of different variants on Total Capture dataset

2D	3D	Hip	Knee	Ankle	Shldr	Elbow	Wrist	Mean (Six)	Others	Mean (All)
SN	PSM	**17.2**	35.7	41.2	50.5	54.8	56.8	37.1	20.3	28.3
ORN	PSM	17.4	29.9	35.2	49.6	44.2	45.1	32.8	20.4	25.4
SN	ORPSM	18.3	25.8	34.0	44.8	44.2	49.8	32.1	19.9	25.5
ORN	ORPSM	18.5	**24.2**	**30.1**	**44.8**	**40.7**	**43.4**	**30.2**	**19.8**	**24.6**

Table 3. MPJPE comparison with SOTAs on Total Capture dataset

Approach	IMU	Temporal	Aligned	Subjects(S1,2,3) W2	Subjects(S1,2,3) A3	Subjects(S1,2,3) FS3	Subjects(S4,5) W2	Subjects(S4,5) A3	Subjects(S4,5) FS3	Mean
PVH[1]				48.3	94.3	122.3	84.3	154.5	168.5	107.3
Malleson[2]	√	√		-	-	65.3	-	64	67	-
VIP[3]	√	√	√	-	-	-	-	-	-	26
LSTM-AE[4]		√		**13.0**	23.0	47.0	21.8	40.9	68.5	34.1
IMUPVH[5]	√	√		19.2	42.3	48.8	24.7	58.8	61.8	42.6
Qiu[6]				19	21	28	32	33	54	29
SN + PSM				14.3	18.7	31.5	25.5	30.5	64.5	28.3
SN + PSM			√	12.7	16.5	28.9	21.7	26	59.5	25.3
ORN + ORPSM	√			14.3	**17.5**	**25.9**	**23.9**	**27.8**	**49.3**	**24.6**
ORN + ORPSM	√		√	12.4	14.6	22	19.6	22.4	41.6	20.6

Code & References

[1] Matthew Trumble, et al. Total capture: 3D human pose estimation fusing video and inertial sensors. In BMVC, pages 1–13, 2017.

[2] Charles Malleson, et al. Real-time full-body motion capture from video and imus. In 3DV, pages 449–457. IEEE, 2017.

[3] Timo von Marcard, et al. Recovering accurate 3d human pose in the wild using imus and a moving camera. In ECCV, pages 601–617, 2018.

[4] Matthew Trumble, et al. Deep autoencoder for combined human pose estimation and body model upscaling. In ECCV, pages 784– 800, 2018.

[5] Andrew Gilbert, et al. Fusing visual and inertial sensors with semantics for 3d human pose estimation. IJCV, 127(4):381–397, 2019.

[6] Haibo Qiu, et al. Cross view fusion for 3d human pose estimation. In ICCV, pages 4342–4351, 2019.