xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera

[1] facebook Reality Labs

Denis Tome [1,2] Patrick Peluse [1] Lourdes Agapito [2] Hernan Badino [1]

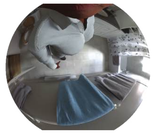
[2] University College London

Problem description

From egocentric input RGB image to 3D joint positions

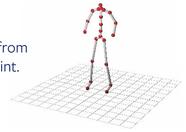
Camera setup

External view of person wearing VR goggles



Input

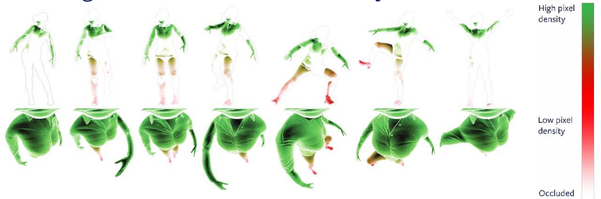
Single RGB image from egocentric viewpoint.



Output

3D skeletal pose

Challenges: self-occlusion and fisheye camera distortion



- Severe self-occlusion
- Strong perspective distortion
- Diminishing pixel density for lower body joints

Existing camera setup for egocentric 3D human pose estimation

EgoCap [38]

Dual fisheye camera system with 2 cameras mounted approx. 15 cm away from the face

Mo2Cap2 [56]

Monocular system with camera mounted on a baseball hat.

Ours

Product oriented solution with fisheye camera embedded in the VR goggles

xR-EgoPose synthetic dataset

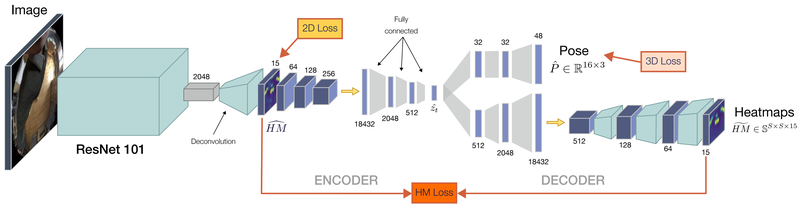
- photo-realistic dataset
- 46 characters
  - 23 female characters
  - 23 male characters
  - 4 body specializations
  - 7 skin colors plus variations
  - 14 clothing items rendered with different colors and textures
- 383.000 frames
  - Train-set: 252.000 frames
  - Test-set: 115.000 frames
  - Validation-set: 16.000 frames
- 9 broad action categories
  - with a total of 64 different actions
- 5 passes (1024 x 1024)
  - RGB
  - Depth
  - Normals
  - Body Segmentation
  - Pixel world position

Frame quality comparison with *mo2cap2 dataset*

- Unrealistic lighting
- Poor image quality
- Unrealistic character textures

Architecture description

Dual-branch autoencoder architecture



Novel **dual-branch** autoencoder architecture, trainable in a **semi-supervised** manner

$$L = \underbrace{||\widehat{\mathbf{HM}} - \mathbf{HM}^{gt}||^2}_{\text{HM Loss}} + \underbrace{\lambda_P(||\hat{\mathbf{P}} - \mathbf{P}^{gt}||^2 + \lambda_\theta \sum_l^L \frac{\mathbf{P}_l \cdot \hat{\mathbf{P}}_l}{||\mathbf{P}_l|| * ||\hat{\mathbf{P}}_l||} + \lambda_L \sum_l^L ||\mathbf{P}_l - \hat{\mathbf{P}}_l||)}_{\text{3D Loss}} + \underbrace{\lambda_{\widetilde{hm}}||\widetilde{\mathbf{HM}} - \widehat{\mathbf{HM}}||^2}_{\text{2D Loss}}$$

1. **Single-branch** learning

$L =$ 2D Loss + 3D Loss

- Strong supervision
- Uncertainty information not encoded

2. **Dual-branch** learning

$L =$ 2D Loss + 3D Loss + HM Loss

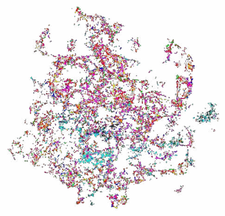
- Strong supervision
- Encodes uncertainty of predicted heatmaps

3. *Weakly supervised* learning

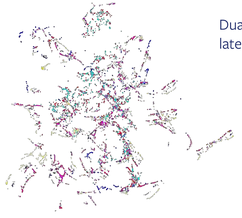
$L =$ 2D Loss + HM Loss

- Weak supervision
- Encodes uncertainty of predicted heatmaps

Dual vs. single branch autoencoder



Single-branch latent space distribution



Dual-branch latent space distribution

Results on xR-EgoPose

Approach	Gaming	Gesticulating	Greeting	Lower Stretching	Patting	Reacting	Talking	Upper Stretching	Walking	**Error (mm)**
Martinez [27]	109.6	105.4	119.3	125.8	93.0	119.7	111.1	124.5	130.5	122.1
Ours - single branch	138.3	108.5	100.3	133.3	117.8	175.6	93.5	129.0	131.9	130.4
Ours - dual branch	56.0	50.2	44.6	51.1	59.4	60.8	43.9	53.9	57.7	**58.2**

Results by Martinez et al. [27] refer to the model trained entirely on xR-EgoPose dataset.
Using the second branch, **only at training time**, brings over 55% improvement compared to the single branch architecture.

Quantitative model evaluation

Human3.6M dataset

Approach	Chen	Moreno	Tome	Zhou	Martinez	Kanazawa	Sun	Fang	**Ours**	Sun
Error (mm)	82.7	76.5	70.7	55.3	47.7	58.8	48.3	45.7	45.2	**40.6**

Evaluation on Human3.6M [19, 7] dataset using evaluation protocol 2 with our novel dual-branch autoencoder architecture

Mo2cap2 dataset

Approach	Indoor dataset error (mm)	Outdoor dataset error (mm)
3DV'17 [28]	76.28	94.46
VCNet [29]	97.85	113.75
Xu [56]	61.40	80.64
Ours	**48.16**	**60.19**

Performance comparison of our model on the mo2cap2 egocentric dataset. Our model has been trained entirely on the mo2cap2 training set without any additional source. Our approach has over 25% **improvement** on both indoor and outdoor test sets.

Weakly supervised training

xR-EgoPose training data

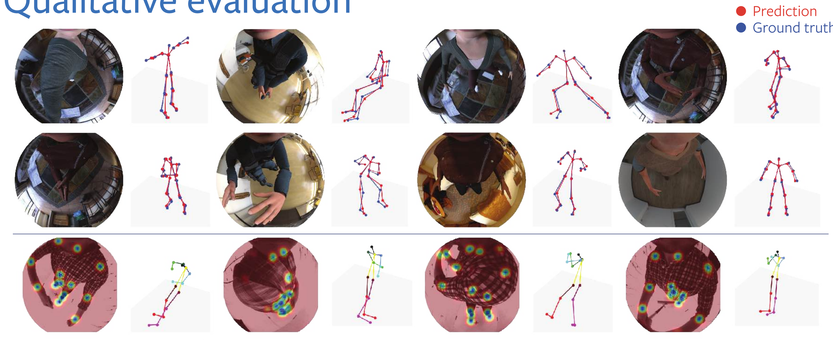
3D	2D	Error (mm)
50 %	50 %	68.04
50 %	**100 %**	**63.98**

Human3.6M training data

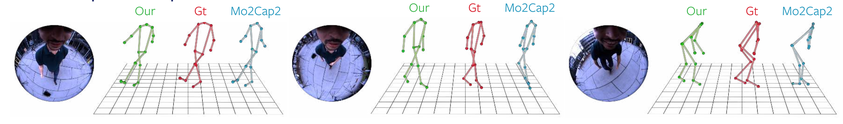
Datasets	Error (mm)
H3.6M	67.9
H3.6M **+ COCO + MPII**	**53.4**

Due to the architecture definition, the model can be trained relying also on existing datasets or augmenting only some labels with better performance results.

Qualitative evaluation



Mo2Cap2 comparison



Acknowledgement

This work was partly funded by the SecondHands project, from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 643950.

[7] C. Ionescu, F. Li. Latent structured models for human pose estimation. In Interna- tional Conference on Computer Vision, 2011
[19] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predic- tive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence, 2014
[27] J. Martinez, R. Hossain, J. Romero, and J. Little. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the International Conference on Computer Vision, 2017
[38] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H. Seidel, B. Schiele, and C. Theobalt. Egocap: egocentric marker-less mo- tion capture with two fisheye cameras. ACM Transactions on Graphics, 2016
[56] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H. Seidel, and C. Theobalt. Mo2Cap2 : Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE Transactions on Visualization and Computer Graphics