SPEC: Seeing People in the Wild with an Estimated Camera

Muhammed Kocabas[1,2], Chun-Hao Paul Huang[1], Joachim Tesch[1], Lea Müller[1], Otmar Hilliges[2], Michael J. Black[1]

[1] MPI for Intelligent Systems, Tübingen, [2] ETH Zurich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

ETH zürich

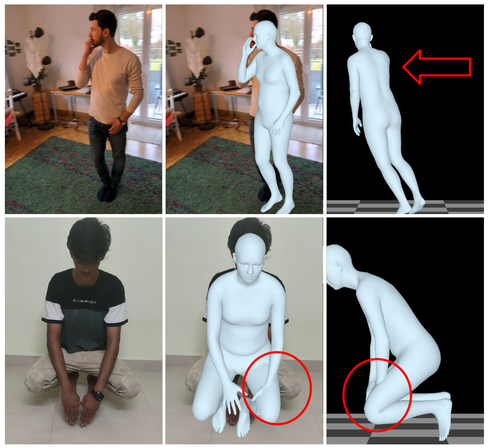
2021 ICCV OCTOBER 11-17 VIRTUAL

Motivation

To reconstruct human bodies given a single image, existing methods assume

1. **Intrinsic:** weak perspective projection - large focal length f=5000
2. **Extrinsic:** no camera rotation w.r.t. the world

These assumptions lead to below errors in real life images

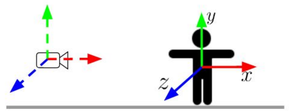


tilted body

wrong body pose

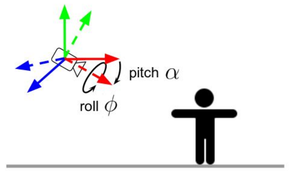
Camera Geometry

To solve this problem, we estimate the camera parameters denoted in (b) from a single image.



$$R^c = \mathrm{I},\ R^b \equiv R^b_c \neq \mathrm{I}$$
$$t^c \neq 0,\ t^b = 0$$

(a) Camera geometry in SOTA methods (IWP-Cam)

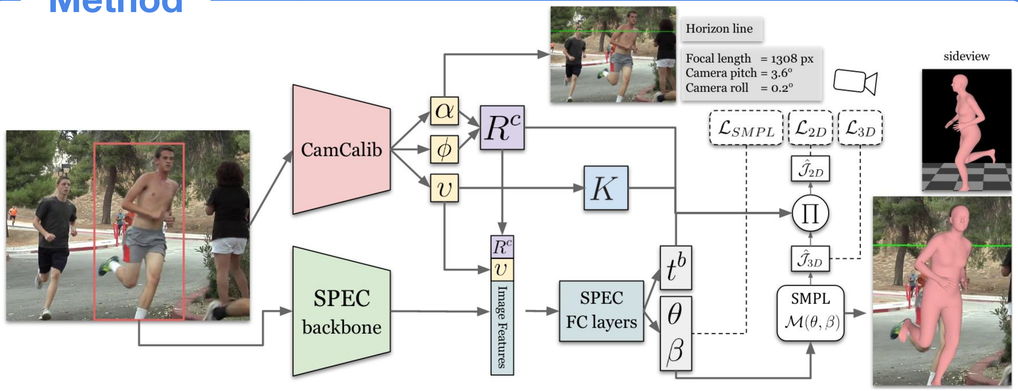


$$R^c \neq \mathrm{I},\ R^b \neq \mathrm{I}$$
$$t^c = 0,\ t^b \neq 0$$

world coord.
camera coord.

(b) Camera geometry of SPEC

Method



- CamCalib takes the whole input image as input and predicts camera parameters. Horizon line (green) shows the predicted camera rotation.
- SPEC takes a cropped bounding box as input and extracts image features using a CNN backbone. Predicted camera parameters from CamCalib are concatenated with image features to estimate SMPL body parameters.
- Camera parameters are also taken into account when computing a loss between the projected 3D joints and ground truth.

Datasets

Pano360 - to train CamCalib

(a) Full panorama image
(b) Random viewpoints

SPEC-MTP

Evaluation only - Subjects are captured from multiple views while mimicking a reference pose. Camera parameters are obtained w.r.t. the global orientation presented in reference poses.

SPEC-SYN

Training and evaluation - Synthetic images (AGORA) - cameras are randomly sampled

Results

Evaluation Metrics

- MPJPE (mean per joint error)), PA-MPJPE are the commonly used metric.
- These metrics exists specifically because current methods reconstruct bodies in camera coordinates.
- Instead, we propose variants of MPJPE and PVE that compute the error in world coordinates without the need of camera information and dub them W-MPJPE and W-PVE.

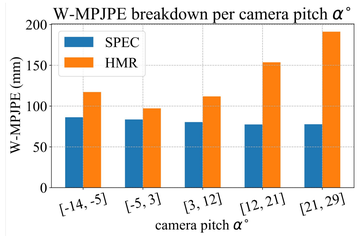
Results on SPEC-MTP

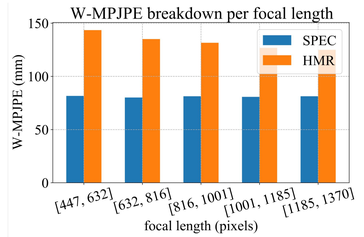
Methods	W-MPJPE	PA-MPJPE	W-PVE
GraphCMR [36]	175.1 / 166.1	94.3	205.5 / 197.3
SPIN [35]	143.8 / 143.6	79.1	165.2 / 165.3
PartialHumans [52]	158.9 / 157.6	98.7	190.1 / 188.9
I2L-MeshNet† [44]	167.2 / 167.0	99.2	199.0 / 198.1
HMR* [28]	142.5 / 128.8	**71.8**	164.6 / 150.7
SPEC	**124.3 / 124.3**	**71.8**	**147.1 / 147.1**

Ablation Experiments

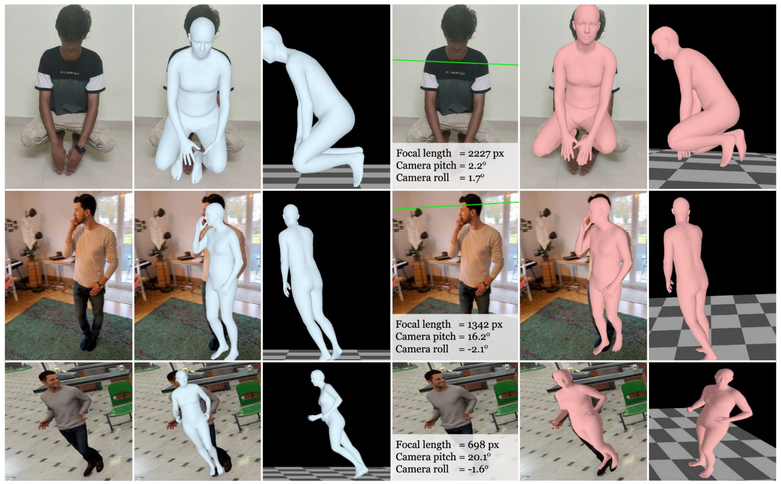
Methods	W-MPJPE	PA-MPJPE	W-PVE
HMR*	128.7 / 96.4	55.9	144.2 / 111.8
HMR* + c	120.4 / 84.2	54.0	135.3 / 98.8
HMR* + $c + f$	118.3 / 85.1	54.0	132.8 / 99.7
HMR* + $c + f + R^c$	77.2 / 77.2	55.3	93.8 / 93.8
SPEC	**74.9 / 74.9**	**54.5**	**90.5 / 90.5**

Error Breakdown w.r.t. the Camera Parameters





Qualitative Results



(a) Input Image (b) HMR* - front (c) HMR* - side (d) CamCalib (e) SPEC - front (e) SPEC - side

ArXiv: https://arxiv.org/abs/xxxxxxx

Project page: spec.is.tue.mpg.de