Expressive Body Capture: 3D Hands, Face, and Body from a Single Image

Georgios Pavlakos[*1,2], Vasileios Choutas[*1], Nima Ghorbani[1], Timo Bolkart[1], Ahmed A. A. Osman[1], Dimitrios Tzionas[1], Michael J. Black[1]

[1]MPI for Intelligent Systems, Tübingen, DE, [2]University of Pennsylvania, PA, USA

[*] Equal contribution
Work performed at MPI

{gpavlakos, vchoutas, nghorbani, tbolkart, aosman, dtzionas, black}@tuebingen.mpg.de

MAX-PLANCK-GESELLSCHAFT

Penn Engineering GRASP LABORATORY

smpl-x.is.tue.mpg.de

Models, Data, & Code

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Goal

Understanding human behavior

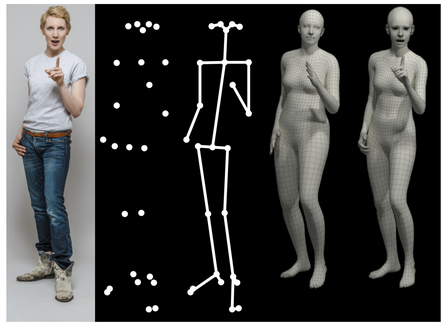
- To understand human behavior we need to capture more than the major body joints; we need:
  - a realistic *body*,
  - an expressive *face*,
  - articulated *hands*.

SMPL [3]	SMPL+H [2]	Frank [5]
✓	✓	✗
✗	✗	✓
✗	✓	✓

- No existing model captures all above factors.

Expressive Body Capture from a single RGB image

1. Create *SMPL-X*, a more expressive human model.
2. Fit *SMPL-X* to the RGB image using:
   (a) Data term with 2D joints from OpenPose [6].
   (b) Automatic gender detection.
   (c) New data-driven DNN pose prior.
   (d) Fast and accurate interpenetration penalty.



Input image | Major joints | Skeleton | SMPL | SMPL-X

SMPL-X

Follows the formulation of SMPL [2]:

$$M(\beta,\theta,\psi) = W(T_p(\beta,\theta,\psi), J(\beta), \theta, \mathcal{W}) \in \mathbb{R}^{10475\times 3}$$

$$T_P(\beta,\theta,\psi) = \bar{T} + B_S(\beta;\mathcal{S}) + B_E(\psi;\mathcal{E}) + B_P(\theta;\mathcal{P})$$

- Body shape: β
- Facial expression: ψ
- Pose: $\theta = [\theta_b, \theta_f, \theta_h]$
  - θ_f: Jaw pose
  - θ_h: Hand pose
  - θ_b: Body pose
- Simple parameterization, efficient, differentiable, compatible with graphics software.

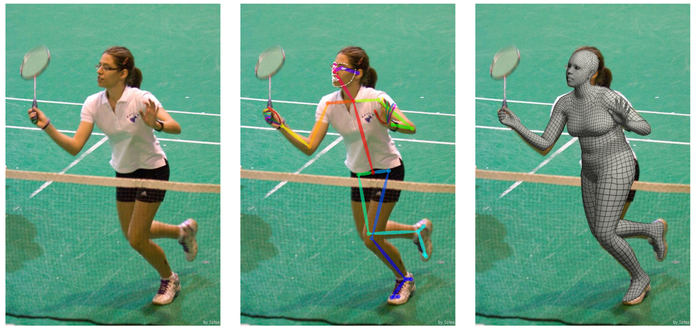
- Joint shape space for body, face and hands.
- Hand pose space of MANO [2]
- Expression space of FLAME [4].
- Separate female, male and neutral models.

EHF Dataset

Expressive Hands and Faces:
- Pseudo ground-truth
- 100 { RGB images, SMPL-X alignments }

Raw scan | Overlay scan & SMPL-X | SMPL-X alignment

SMPLify-X

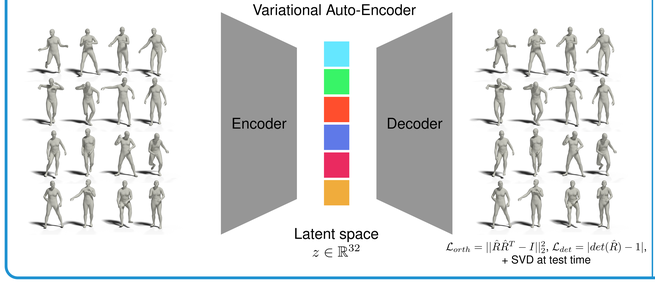


Input image | Keypoint detection | Model fitting

Bottom-up **Top-down**

$$E(\theta,\beta,\psi) = E_J(\beta,\theta;K,J_{est}) \quad \text{Data term (2D Joints reprojection)}$$
$$+\lambda_{\theta_b} E_{\theta_b}(\theta_b) + \lambda_{\theta_f} E_{\theta_f}(\theta_f) + \lambda_{\theta_h} E_{\theta_h}(\theta_h) + \lambda_{\theta_\alpha} E_{\theta_\alpha}(\theta_b) \quad \text{Pose priors}$$
$$+\lambda_\beta E_\beta(\beta) + \lambda_\mathcal{E} E_\mathcal{E}(\psi) \quad \text{Shape and expression priors}$$
$$+\lambda_\mathcal{C} E_\mathcal{C}(\theta,\beta,\psi) \quad \text{Interpenetration penalty}$$

VPoser



Body pose prior loss

$$E_{\theta_b}(\theta_b) = \|\theta_b - Decoder(Encoder(\theta_b))\|_2^2$$

$$\theta_b = Decoder(z)$$

Gender Classifier

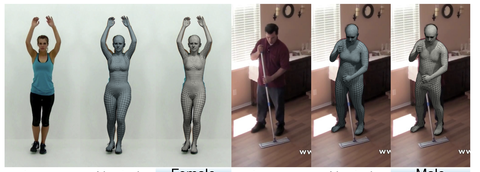
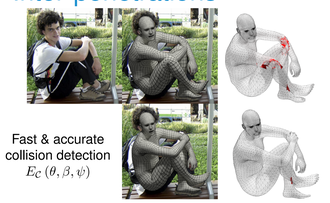


Image | Neutral model | Female model | Image | Neutral model | Male model

Inter-penetrations



Fast & accurate collision detection
$E_\mathcal{C}(\theta,\beta,\psi)$

Experiments

Increasing expressiveness

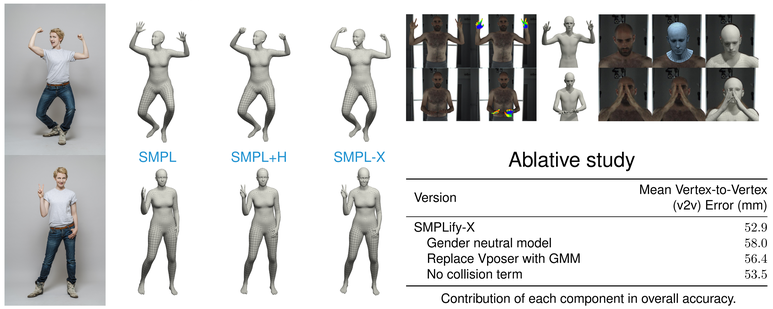
Model	Keypoints	Mean Vertex-to-Vertex (v2v) error (mm)	Joint error (mm)
"SMPL"	Body	60.4	66.6
"SMPL"	Body+Hands+Face	66.4	86.5
SMPL-X	Body+Hands+Face	**55.7**	**65.4**

Richer modeling power results in lower errors.

Part-based vs holistic model

Method	v2v (mm), head only	MPJPE (mm), hands only
Panteleris et al [7]	-	26.5
FLAME	6.46	–
SMPLify-X	**5.32**	**19.8**

The holistic model performs better than part-based models.



SMPL | SMPL+H | SMPL-X

Ablative study

Version	Mean Vertex-to-Vertex (v2v) Error (mm)
SMPLify-X	52.9
Gender neutral model	58.0
Replace Vposer with GMM	56.4
No collision term	53.5

Contribution of each component in overall accuracy.

Qualitative

SMPL-X vs Frank

RGB | TotalCapture [5]: ~ 500 cameras | SMPL-X fitting: ~ 500 cameras | SMPL-X fitting: 1 camera

References

1. Bogo et al. *Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image.* ECCV, 2016.
2. Romero et al. Embodied hands: *Modeling and capturing hands and bodies together.* TOG, 2017.
3. Loper et al. *SMPL: A skinned multi-person linear model.* TOG, 2015.
4. Li et al. *Learning a model of facial shape and expression from 4D scans.* TOG, 2017.
5. Joo et al. *Total capture: A 3D deformation model for tracking faces, hands, and bodies.* CVPR, 2018.
6. Cao et al. *OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields.* arXiv, 2019.
7. Panteleris et al. *Using a single RGB frame for real time 3D hand pose estimation in the wild.* WACV, 2018.