Scan me for code and data!

forte@is.mpg.de

Reconstructing Signing Avatars From Video Using Linguistic Priors

Maria-Paola Forte, Peter Kulits, Chun-Hao Paul Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black
Max Planck Institute for Intelligent Systems, Stuttgart and Tübingen, Germany

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

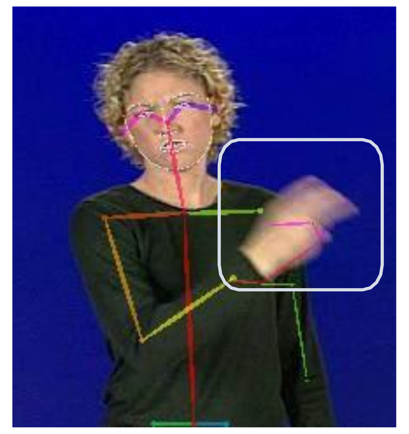
Goal

Reconstruct expressive 3D avatars from monocular sign-language (SL) video of isolated signs.

Input frame

Problem

Current pose-estimation methods struggle with SL video due to difficult hand occlusions and motion blur caused by the fast hand movements that are typical of SL.



Keypoint detection

Model fitting with SMPLify-SL
(SMPLify-X[1] for SL)

Key Idea

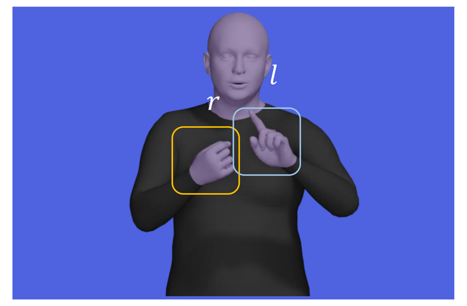
Leverage linguistic rules of SL to develop novel priors that help disambiguate hand poses in SL videos.

Model fitting with **SGNify**

Method

SGNify is based on the expressive SMPLify-X[1] body-reconstruction method.

Input frame



SMPLify-SL

We introduce two novel linguistic priors to increase the image evidence for a pose and improve hand-pose estimates for challenging videos:

- *Hand-Pose Symmetry:* penalizes differences between the right and left hand-pose estimates

$$L_s = \lambda_s \|\theta_t^r - r(\theta_t^l)\|_2^2$$

- *Hand-Pose Invariance:* penalizes differences between the reference hand pose and the estimated hand pose. Each sign is defined by a characteristic reference pose sequence (θ^h_{ref}) which defines the hand pose that we expect at each time t

$$L_h = \lambda_h \|\theta^r_{ref,t} - \boxed{\theta^r_t}\|_2^2 \quad \text{OR} \quad \lambda_h \|\theta^l_{ref,t} - \boxed{\theta^l_t}\|_2^2$$

$\lambda_h \,\|$  $-$  $\|_2^2$

The hand pose is either *static* throughout the sign articulation or *transitions* from one pose to another pose.

We introduce eight *sign-group classes* and apply the linguistic constraints that are active for the detected class.

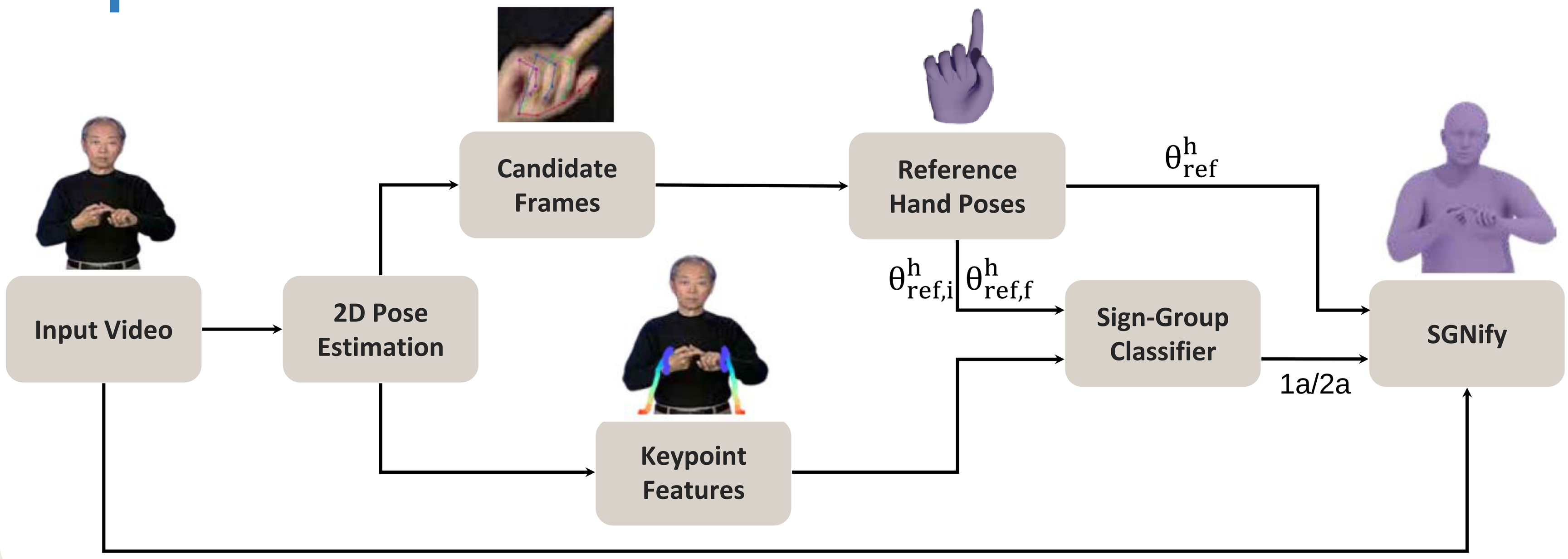
Class	Symmetry	Invariance: Dominant	Invariance: Non-dominant
0a	✗	static	✗
1a	✓	static	static
2a	✓	static	static
3a	✓	static	static

Class	Symmetry	Invariance: Dominant	Invariance: Non-dominant
0b	✗	transitioning	✗
1b	✓	transitioning	transitioning
2b	✗	transitioning	static
3b	✗	transitioning	static

Dataset

We collected isolated signs and sentences paired with ground-truth MoCap meshes.

Pipeline



Results

Quantitative

Mean Vertex-to-Vertex Error (mm) of the 57 German signs in our dataset

Method	Upper Body	Left Hand	Right Hand
PIXIE[2]	60.11	25.02	22.42
PyMAF-X[3]	68.61	21.46	19.19
SMPLify-SL	56.07	22.23	18.83
OSX[4]	**52.62**	38.90	37.58
SGNify	55.63	**19.22**	**17.50**

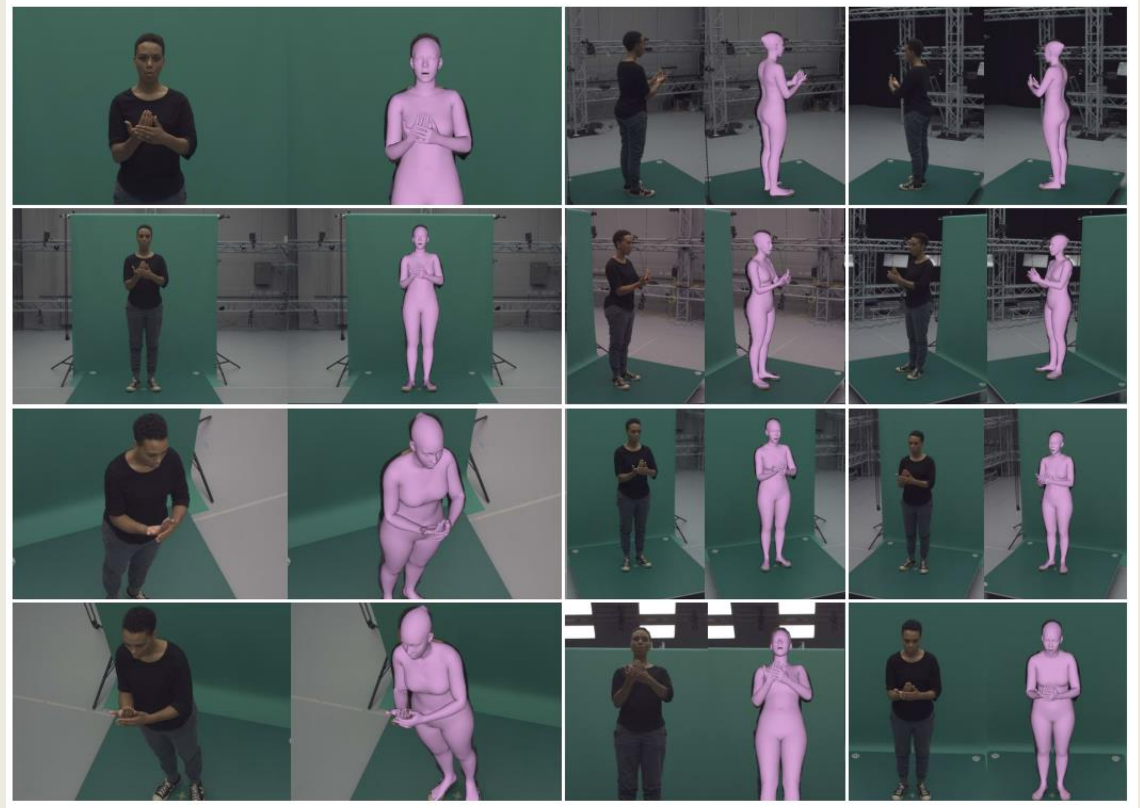
Perceptual Study

20 proficient signers evaluated 50 American signs

Method	Average Recognition Rate (%)
Real Video	90.9
PyMAF-X[3]	62.0
SMPLify-SL	74.8
SGNify	**86.2**

Extensions

Multi-View



Continuous Signing

Der **Vater** muss für die Reparatur seines **Autos** viel Geld **ausgeben**.

imprs-is
Acknowledgments: We thank the Perceiving Systems Data Team and the MPI-IS Optics and Sensing Lab for their support in capturing the dataset.

[1] Pavlakos et al., SMPLify-X, CVPR 2019 [2] Feng et al., PIXIE, 3DV 2021 [3] Zhang et al., PyMAF-X, PAMI 2023 [4] Lin et al., OSX, CVPR 2023

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS