3D Hands, Face and Body Extraction for Sign Language Recognition

Agelos Kratimenos[1]
ageloskrat@yahoo.gr

Georgios Pavlakos[2]
pavlakos@seas.upenn.edu

Petros Maragos[1]
maragos@cs.ntua.gr

[1]School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece

[2]Dept. of Computer and Information Science
University of Pennsylvania
Philadelphia, USA

SLRT 2020

Sign Language Recognition, Translation & Production

Technical Approach

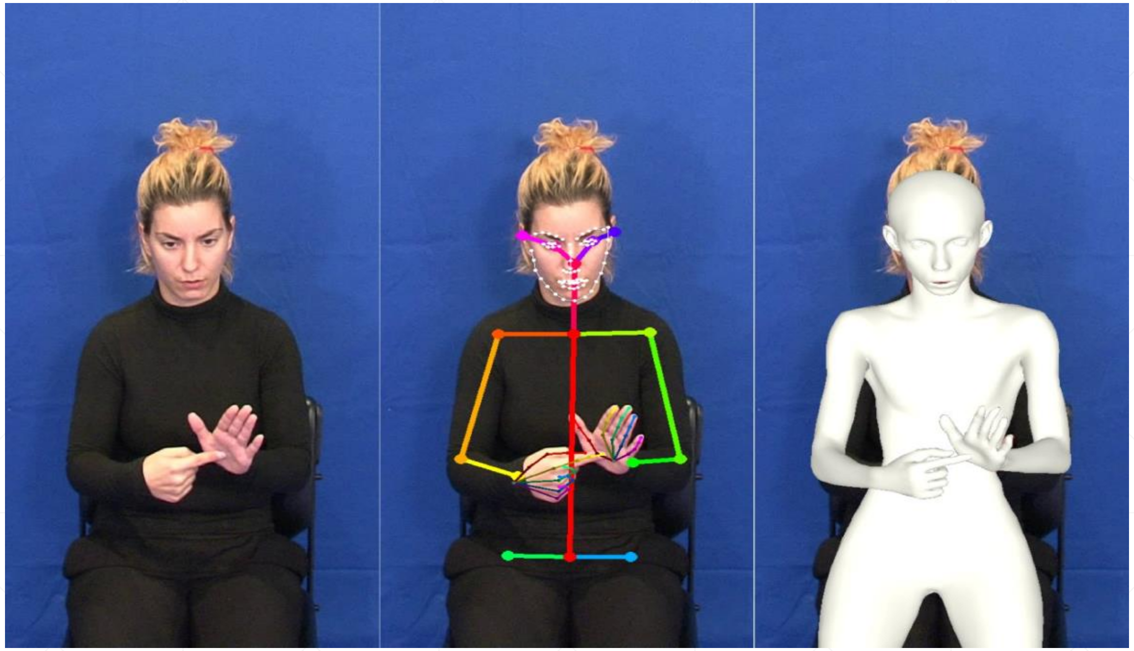
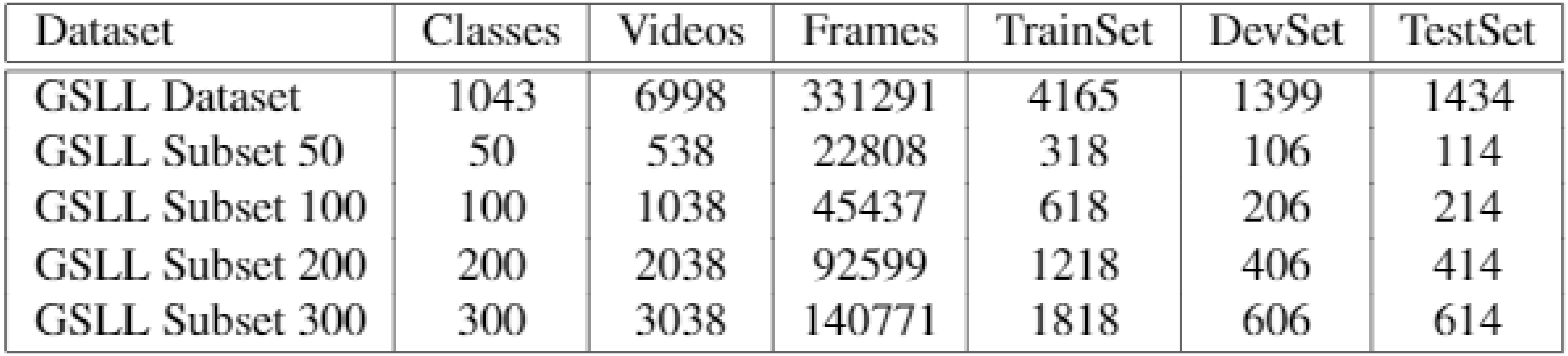


Figure 1: Raw image of a single frame from the sequence of the sign X, the skeleton of the image produced by Openpose and the 3D body reconstruction of it produced by SMPL-X.

- **Comparison of three methods for Independent Sign Language Recognition task:**
  - **Raw Images**: Feed a Conv3D-LSTM network with sequences of 175x175 frames.
  - **Openpose**: Extract a 411-parameters 2D skeleton for each frame and feed each frame sequence into an LSTM network.
  - **SMPL-X**: Extract a 3D body, face and hand reconstruction of 88 features for each frame and feed each frame sequence into an LSTM network.

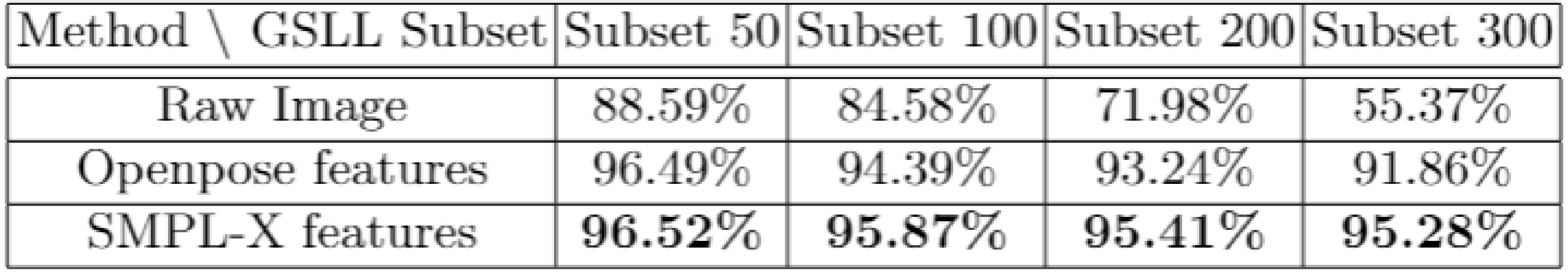


Dataset	Classes	Videos	Frames	TrainSet	DevSet	TestSet
GSLL Dataset	1043	6998	331291	4165	1399	1434
GSLL Subset 50	50	538	22808	318	106	114
GSLL Subset 100	100	1038	45437	618	206	214
GSLL Subset 200	200	2038	92599	1218	406	414
GSLL Subset 300	300	3038	140771	1818	606	614

Table 1: Statistics for the Greek Sign Language Lemmas Dataset and its respective subsets.

- **Evaluation on the Greek Sign Language Lemmas Dataset (GSLL) :**
  - Consists of two signers
  - Each sign is repeated from 5 to 17 times.
  - Contains 1043 different classes (signs).

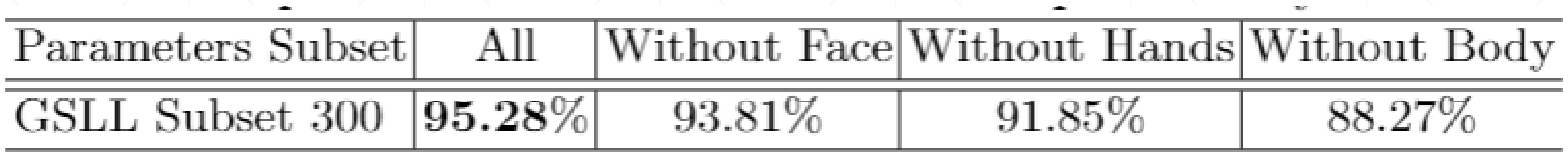
Results and Discussion



Method \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300
Raw Image	88.59%	84.58%	71.98%	55.37%
Openpose features	96.49%	94.39%	93.24%	91.86%
SMPL-X features	**96.52%**	**95.87%**	**95.41%**	**95.28%**

Table 2: Comparison of the three representations for sign classification: i) Raw RGB images ii) Openpose 2D skeleton keypoints and iii) SMPL-X parameters.

- **Main experiment:**
  - **Conv3D-LSTM** model (43 million parameters) declines significantly in performance with the increase in different signs.
  - **Openpose** (1.4 million parameters) and SMPL-X (0.7 million parameters) outperform the convolutional model since they manage to eliminate the redundant information from each frame.
  - **SMPL-X** seems to outperform Openpose especially with the increase in the number of different signs dictating that a more detailed and qualitative representation of the human body is needed for the SLR task. While varying and more complex signs are being added to the train set, Openpose fails to convey the small details that differentiate these signs, while SMPL-X holds its accuracy almost fixed.



Parameters Subset	All	Without Face	Without Hands	Without Body
GSLL Subset 300	**95.28%**	93.81%	91.85%	88.27%

Table 3: Experiments with subset of features provided by SMPL-X.

- **Ablation Study:**
  - We experiment with different subsets of SMPL-X parameters to point out the importance of combining all three channels of information (body, face and hands) for SLR task.
  - We train the LSTM network once without the facial expression parameters (jaw pose, left and right eye pose, expression) (69 features), once without hand keypoints (64 features) and once without body information (50 features).
  - Indeed, omitting any of these three channels reduces the accuracy in the GSLL Subset emphasizing on the importance of finding a qualitative method to combine them together.
  - Interestingly, body seems more important than hands which can be attributed to the fact that when few and simple signs are available, the sign can be mainly conveyed through the movement of the arms while the hands are commonly remain straight.

Thank you for your attention!