Independent Sign Language Recognition with 3D Body, Hands, and Face Reconstruction

Agelos Kratimenos[1,3], Georgios Pavlakos[2], and Petros Maragos[1,3]

[1] School of ECE, National Technical University of Athens / [2] School of EECS, University of California Berkeley / [3] Robot Perception and Interaction Unit, Athena Research Center

Paper #3776

1. Outline

- **Isolated Sign Language Recognition (SLR)**: The task in which one wants to recognize the sign performed by a signer in a video.
- **Until now**: State-of-the-art works have managed to deeply elaborate on these features independently, but no work has adequately combined all three channels of information, namely, face, body and hands.

2a. SMPL-X and SMPLify-X

- **SMPL-X**: a contemporary parametric model that enables joint extraction of 3D body shape, face and hands information from a single image.
- It is a qualitative way to combine face, hands and body information, with great detail which is absolutely needed in the task of SLR.
- **SMPLify-X**: estimates 2D parameters, using OpenPose, and then optimizes model parameters to fit the features; a procedure which takes up to a minute for a single frame.
- SMPL-X produces a total of **88 parameters**: 10 for shape parameters, 3 for global orientation, 24 for left and right hand pose, 3 for jaw pose, 6 for left and right eye pose, 10 for expression and 32 for the body pose.

2b. Dataset

- The Greek Sign Language Lemmas Dataset (GSLL) consists of two signers, signing in almost 3500 videos.
- The dataset consists of 347 different signs or classes.

GSLL Subset	Videos	Frames	TrainSet	DevSet	TestSet
50 classes	538	22808	318	106	114
100 classes	1038	45437	618	206	214
200 classes	2038	92599	1218	406	414
300 classes	3038	140771	1818	606	614
347 classes	3464	161050	2066	695	703

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

2c. Features

- **Openpose**: We extract 411 parameters for each frame and feed the sequence in an RNN consisting of one Bi-LSTM layer of 256 units and a Dense layer for classifying, after applying standard scaling to our features. We believe that by providing a recurrent network with these features will eliminate any redundant information (e.g background, clothes, lighting) that a raw image contains.
- **Raw Image and Optical flow**: A 3D state-of-the-art method for action recognition and signing is the I3D network. The 3D convolution module used, exploits both raw RGB frames and the optical flow of these. We reshape each frame to a 175×175 array and normalize its pixels to $[0, 1]$.
- **SMPL-X**: Due to its ability to interpret the structure of the body in detail, we strongly believe that this method will provide key features for this task. Moreover, SMPL-X provides 3D information, in comparison to Openpose that results to 2D only keypoints, so the extracted features should be strictly more informative. This method extracts 88 features per frame, creating a (length of sequence) $\times$ 88 array for each sequence, which is being standard scaled as in the Openpose experiments. Similarly to Openpose, we employ the same neural network architecture so that we can directly compare the two methods independently of the type of architecture.

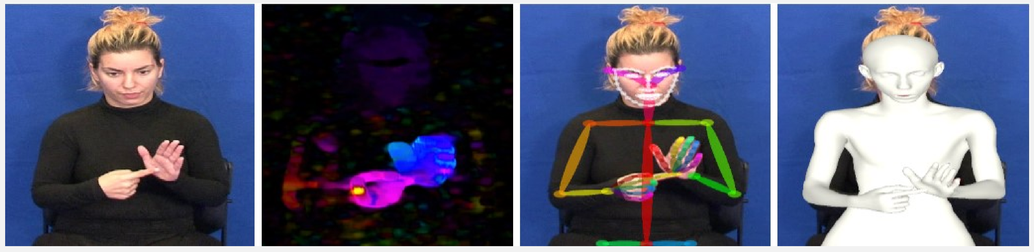

Figure: **i) First image: Raw RGB frame, ii) Second Image: Optical flow of a frame, iii) Third Image: Openpose 2D Skeleton, iv) Fourth image: 3D Body Reconstruction produced by SMPL-X.**

4. Results

- The next table shows the results from the three methods described before, for a different amount of classes:

Method \ GSLL Subset	Subset 50	Subset 100	Subset 200	Subset 300	Full Dataset	Parameters
3D RGB & Optical Flow Images	90.41%	86.85%	80.79%	71.36%	65.95%	43.41 million
2D Openpose Skeleton	96.49%	94.39%	93.24%	91.86%	88.59%	1.55 million
3D SMPL-X Reconstruction	**96.52%**	**95.87%**	**95.41%**	**95.28%**	**94.77%**	0.88 million

Table: **Comparison of the three methods for training: i) Raw RGB images and their Optical Flow ii) Openpose skeleton key-points and iii) 3D Body Reconstruction key-points.**

4 Results

- The convolutional model consists of almost 50 million parameters while the Openpose RNN model and the SMPL-X one, consist of around 1 million parameters.
- Despite the fact that the 3D Convolutional model starts very good at around 90%, it quickly drops to 65%, when there are more classes. This is typical for convolutional models.
- Both 2D Openpose skeleton features and SMPL-X ones keep their accuracy almost steady, with the latter losing only 2% from 50 to 347 classes.
- While Openpose is good too, it quickly diverges from its 96.5% starting point.

Ablation Study:

Parameters	Openpose	SMPL-X
All	**88.59%**	**94.77%**
Without Face	88.34%	93.19%
Without Hands	70.20%	89.58%
Without Body	84.21%	85.02%

Table: **Experiments with subset of features produced by Openpose and SMPL-X.**

- All 3 channels of information matter for this task.
- While the face plays the smallest role, body and specifically the arms matter a lot.

5. Contributions

- We exploited one of the most contemporary methods for reconstructing 3D body, face and hands; SMPL-X.
- We published the Greek Sign Language Lemmas Dataset (GSLL), for further experimentation
- We compared state-of-the-art methods: I3D-type convolutional model with raw images and optical flow, 2D Openpose skeleton and 3D body, face and hands reconstruction.
- We conducted an ablation study to show the importance of having all three channels of information for the task of isolated Sign Language Recognition.