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, 15773 Athens, Greece
[2] School of EECS University of California Berkeley, California, USA
[3] Robot Perception and Interaction Unit, Athena Research Center, 15125 Maroussi, Greece
ageloskrat@yahoo.gr, pavlakos@berkeley.edu, maragos@cs.ntua.gr

List of Contents

1. Introduction
2. SMPL-X
3. Method
4. Results & Discussion
5. Contributions

List of Contents

1. Introduction
2. SMPL-X
3. Method
4. Results & Discussion
5. Contributions

Isolated Sign Language Recognition

Isolated Sign Language Recognition (SLR): The task in which one wants to recognize the isolated sign performed by a signer in a video.

- complex visual recognition problem
- combines several challenging tasks of Computer Vision
- exploits and fuses information from hand gestures, body features and facial expressions

Until now: State-of-the-art works have managed to deeply elaborate on these features independently and not in combination.

We aim:

- to adequately combine all three information channels to efficiently recognize Sign Language.
- to exploit a robust and qualitative 3D way, to reconstruct body, face and hand information from a signer.

Propose

What we propose?

- A robust way to combine all three channels of 3D information, namely hands, face and body structure.
- A method that captures every detail in hands and facial expressions.

This will be done with the help of SMPL-X.

List of Contents

1. Introduction
2. SMPL-X
3. Method
4. Results & Discussion
5. Contributions

3D Reconstruction: SMPL-X

SMPL-X [1]: a contemporary parametric model that enables joint extraction of 3D body shape, face and hands information from a single image.

- a qualitative way to combine face, hands and body information
- great detail in hands and face which is absolutely needed in the task of SLR

Figure: Qualitative results of SMPL-X for in-the-wild images.

[1] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas and M. J. Black "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image", CVPR 2019

3D Reconstruction: SMPLify-X

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

List of Contents

1 Introduction

2 SMPL-X

3 Method

4 Results & Discussion

5 Contributions

Greek Sign Language Lemmas Dataset

The main core of our experiments is conducted on the **Greek Sign Language Lemmas Dataset** (GSLL); statistics of which are shown in the next Table. GSLL is an isolated SLR dataset consisting of two signers performing with a uniform background.

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

Preliminary experiments [2] using naive Conv3D network, 2D Openpose features and SMPL-X features revealed interesting results and drawn great attention.

[2] Kratimenos A. et al "3D hands, face and body extraction for sign language recognition", ECCV Workshop 2020

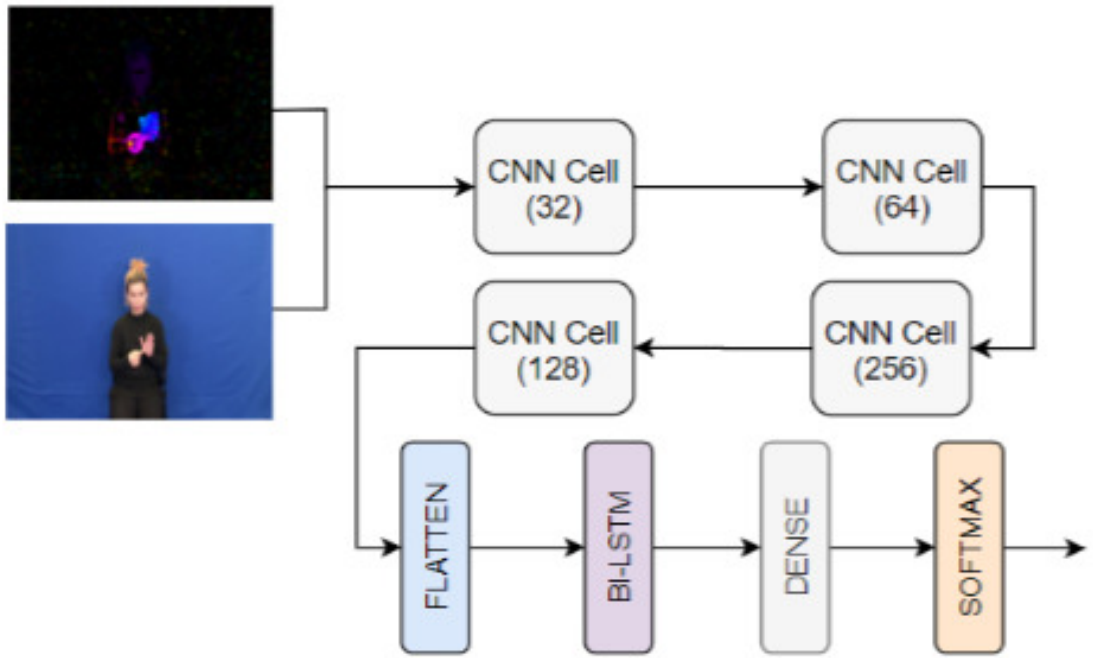
Technical Work

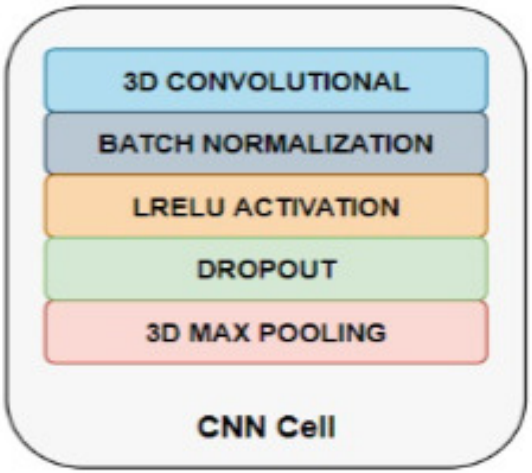
Next,we present the methods with which we confront this problem:

- **Openpose 2D skeleton**: We extract 411 parameters for each frame and feed the sequence in an RNN consisting of one Bi-LSTM layer of 256units and a Dense layer for classifying, after applying standard scaling to our features.
- **Raw Image and Optical flow**: We use a 3D convolutional neural network followed by a Bi-LSTM layer using both RGB frames and optical flow. We also train a VGG16-LSTM model which is initialized with Imagenet weights.
- **3D SMPL-X features**: 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.

The 3D CNN Architecture

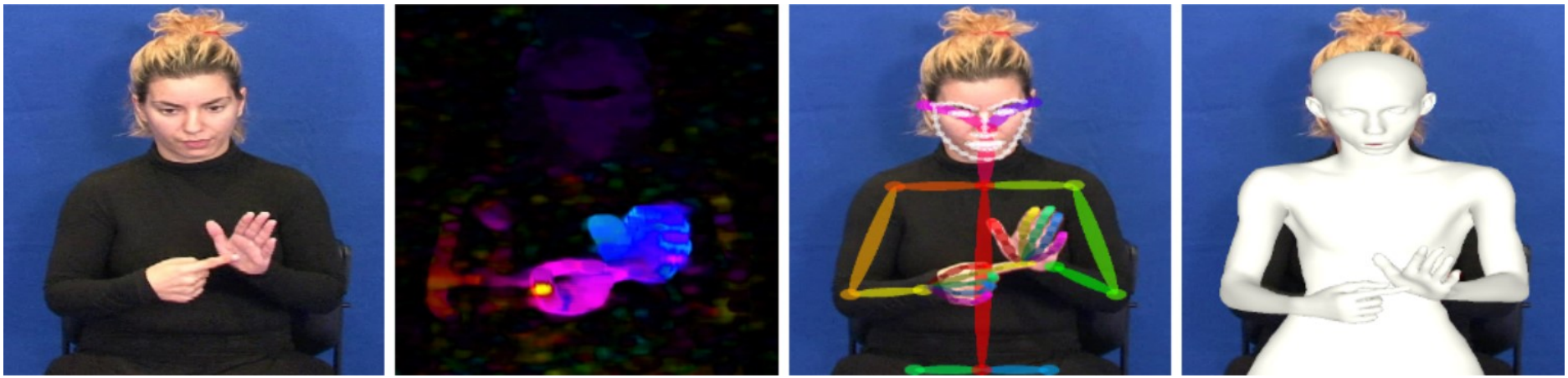
The **3D CNN-LSTM architecture** exploited using both raw frames from the signing video and the optical flow produced by the motion of the hands, the body and the face of the signer.



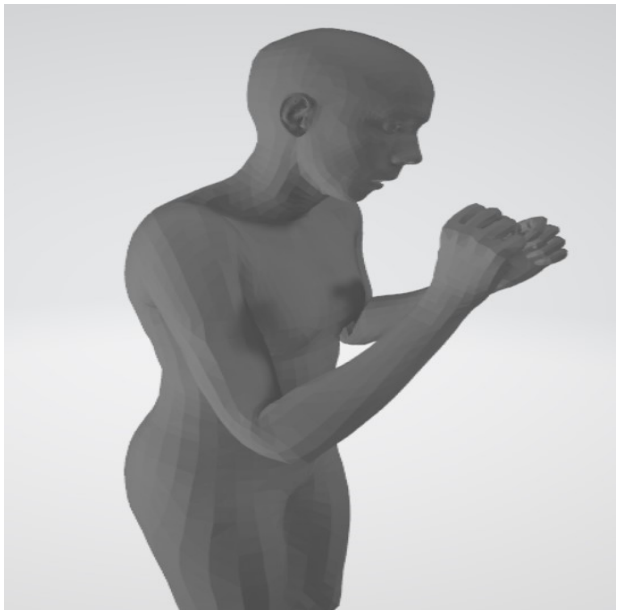


Comparison of features

The features used in the experiments. The first image depicts a raw RGB frame, the second Image shows the optical flow of a frame, the third Image shows the Openpose 2D Skeleton of the signer while the fourth image depicts 3D Body Reconstruction produced by SMPL-X.



3D Reconstruction Example



List of Contents

1 Introduction

2 SMPL-X

3 Method

4 Results & Discussion

5 Contributions

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.
- SMPL-X keeps its accuracy almost steady, losing only 2% from 50 to 347 classes.
- While Openpose is good too, it quickly diverges from its 96.5% starting point.

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

Ablation Study

To further examine the features produced by SMPL-X, we experiment with a combination of a subset of features produced by it.

We remove all information from facial expressions (jaw, left and right eye pose and expression) and train the model again with 69 features. We then only remove the body pose information and train with 50 features. Finally, we remove left and right hand pose and train with a total of 64 parameters.

We conduct the same experiments for Openpose by separating pose keypoints (75 parameters), face keypoints (210 parameters), and left and right hand keypoints (126 parameters).

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%

List of Contents

1. Introduction
2. SMPL-X
3. Method
4. Results & Discussion
5. Contributions

Contributions & Future Work

- Exploited one of the most contemporary methods for 3D body, face and hands reconstruction, namely SMPL-X.
- We publish the GSLL dataset, 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.

Thank you for your attention!