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

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

[1] School of Electrical & Computer Engineering
National Technical University of Athens, Greece
[2] Department of Computer and Information Science
University of Pennsylvania, Philadelphia, USA

Abstract. For the problem of Sign Language Recognition (SLR), the majority of the information is included in three main channels; hand gestures, facial expression and body pose. While many state-of-the-art works have managed to deeply elaborate on these features independently, to the best of our knowledge, no work has adequately combined all these three information channels, particularly in 3D, to efficiently recognize Sign Language. In this work, we employ SMPL-X, a contemporary parametric model that enables joint extraction of 3D body shape, face and hands information from a single image. We use this holistic 3D reconstruction for SLR, demonstrating that it leads to higher accuracy than recognition from raw RGB images, or 2D skeletons. Simultaneously, we demonstrate the importance of combining the information from all three channels, to achieve the best recognition results.

Keywords: 3D body reconstruction, Independent Sign Language Recognition, Greek Sign Language, Facial Expression, Gesture Recognition

1 Introduction

Sign Language Recognition (SLR) is a very hard task due to the need of combining information from three different channels; face, body and hands. Each independent task has already been successful in the past.[5] But combining all these features in the requisite detail for SLR is far from being perfected. We experiment with SMPL-X, a sophisticated tool that can reconstruct with notable precision the human body from a single RGB image. To the best of our knowledge, this is the first work that combines all three channels of information in a qualitative way for SLR. In this paper, we compare i) raw images ii) skeleton reconstruction with Openpose and iii) 3D body reconstruction with SMPL-X, to evaluate the efficiency of our proposed method. A secondary contribution of this work is the study of the connection between the three channels of information and their importance in SLR. We conduct experiments to show that the best results are achieved after all three channels of information are being included.

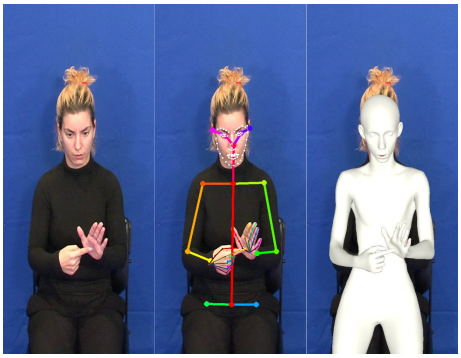

Fig. 1: Left: Raw RGB image, Middle: OpenPose skeleton, Right: 3D body reconstruction produced by SMPL-X.

2 Technical Approach

SMPL-X Model and SMPLify-X reconstruction: SMPL-X is based on a template mesh which has $N = 10475$ vertices, and a kinematic tree with 54 joints. This model is controlled by 162 pose parameters θ which correspond to the 3D rotation of each model joint, shape parameters β and the expression parameters ψ, which correspond to 10 coefficients of a PCA shape space and expression space respectively.

SMPLify-X [4] is a two stage method, responsible for efficiently and effectively reconstructing humans from image data. In the first stage, a set of features is detected on the image, typically 2D keypoints detected by Openpose [1], which include body and hand keypoints and facial landmarks. We refer to these detected 2D keypoints with J_{est}, and we assume that each keypoint $J_{est,i}$ is detected with confidence ω_i. For the second stage, we fit SMPL-X to these 2D landmarks, encouraging the projection of the 3D joints of the model to agree with the detected 2D locations. The complete objective also includes a penalty for mesh intersections and a set of priors for regularization.

Dataset: Since our goal is to test the ability of the proposed method to adequately extract 3D hand, face and body features, we limit our approach to non-continuous sign language recognition. Instead, we focus on the Greek Sign Language Lemmas Dataset (GSLL) [6] which proved to be ideal for our experiments since it consists of two signers and 1043 different signs (classes). We limit our experiments to smaller subsets that include the 50, 100, 200 and 300 most frequent classes respectively in a total of 500 to 3000 videos.

Training Methodology: We do not intervene on the length of each feature sequence, resulting in various lengths from 10 to 300 frames per sign. Next, we present the methods with which we confront this problem (Figure 1).

Raw Image: We reshape each frame in a 175×175 array and normalize its pixels to $[0, 1]$. We feed our images' sequence in a Conv3D-LSTM model,

Table 1: Comparison of the three representations for sign classification: i) Raw RGB images ii) Openpose 2D skeleton keypoints and iii) SMPL-X parameters.

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%**

the structure of which is similar to [3], alongside with a VGG16-LSTM model which is initialized with Imagenet weights and is followed by a Global Average 2D Pooling layer. **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. **SMPL-X**: Due to SMPL-X ability to interpret the 3D structure of the body in detail, we strongly believe that this method will provide key features for action and sign recognition. In this case, we extract 88 features per frame and follow the same procedure as with the Openpose.

We train all three networks using categorical cross-entropy loss. SGD is used to optimize the loss function, with an initial learning rate of 0.0001 and 10% decay rate per epoch, while the batch size is set to 1, due to varying sequence length. We perform Learning Rate Reduction and Early Stopping by monitoring the validation loss with a patience of 3 and 5 epochs respectively.

3 Results and Discussion

According to Table 1, Openpose and SMPL-X models, which consist of 1.4 and 0.7 million parameters respectively, outperform the Conv3D-LSTM and VGG16-LSTM model, which consist of 43 and 15 million parameters respectively. This can be attributed to the fact that the former two eliminate the redundant information from each frame, keeping only the essential key-points. VGG16, in specific, fails to converge and reduce its loss achieving an accuracy below 10% for all classes. This does not come as a surprise to us since Joze and Koller in [2] have trained a VGG16-LSTM model for the MS-ASL dataset which achieved 13.33% for the ASL100 Subset and just 1.47% for the ASL500 Subset. GSLL Dataset is characterised by a very uniform environment between each sign and each signer (only two signers in front of a blue cloth). The MS-ASL dataset for instance, consists of 222 distinct signers where each signer performs in a completely altered environment. We strongly believe that Openpose and mainly SMPL-X will by far outperform convolutional models in these datasets, which simulate more accurately the real world. Finally, SMPL-X seems to outperform the features produced by Openpose especially with the increase of different signs, dictating that a more detailed and representation of the human body is needed for the Sign Language Recognition 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.

Table 2: Experiments with subset of features provided by SMPL-X.

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

To further examine SMPL-X features and the dependence between the three channels of information we experiment with subsets of the total of 88 features. First, we remove facial expression parameters (jaw pose, left and right eye pose and expression) and train the model with a total of 69 features. Secondly, we remove body pose information and train the model with 50 features. Finally, we remove left and right hand key-points and train the model with 64 parameters. Table 2 sums up the results from all the aforementioned experiments.

First of all, we can see that omitting any of these three channels indeed reduces the accuracy of our model. In fact, we expect the omission of facial characteristics to affect even more the accuracy in the continuous sign language where the face plays a crucial role into expressing the intensity of a word. For example, "rain" and "snow" have the exact same hand configurations, whereas only the mouth shape changes. Furthermore, we observe that removing hand information is less harmful than removing body pose. That 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. Nonetheless, both hands and body structure (chiefly due to arms) are of vital importance for SLR while at the same time, omitting facial expression affect the model's optimality.

Concluding, in this paper we investigated the extraction of 3D body pose, face and hand features for the task of Sign Language Recognition. We compared these key-points, to Openpose features, the most famous method for extracting skeleton parameters and features from raw RGB frames. The experiments revealed the superiority of SMPL-X key-points due to the detailed and qualitative features extraction in the three aforementioned regions of interest. Moreover, we exploited SMPL-X to point out the significance of combining all these three regions for optimal results in SLR.

References

1. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. PAMI (2019)
2. Joze, H.R.V., Koller, O.: MS-ASL: A large-scale data set and benchmark for understanding american sign language. BMVC (2019)
3. Kratimenos A, A.K., et al.: Augmentation methods on monophonic audio for instrument classification in polyphonic music. EUSIPCO (2020)
4. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, et al.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019)
5. Pitsikalis, V., et al.: Advances in Phonetics-based Sub-Unit Modeling for Transcription Alignment and Sign Language Recognition. In: CVPR Workshop (2011)
6. Theodorakis, S., et al.: Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. Image and Vision Computing (2014)