Can Everybody Sign Now?
Exploring Sign Language Video Generation
from 2D Poses

Lucas Ventura[1], Amanda Duarte[1,2], and Xavier Giró-i-Nieto[1,2]

[1] Universitat Politècnica de Catalunya, Barcelona, Catalonia/Spain
[2] Barcelona Supercomputing Center, Spain
lucas.ventura.ripol@estudiantat.upc.edu
{amanda.duarte,xavier.giro}@upc.edu

1 Introduction

Sign Language is the primary means of communication of the Deaf community but barely known by the rest of the population. This situation creates difficulties in conversations between sign and non-sign language speakers, which are normally addressed with textual transcriptions of the spoken language, or the sign-speakers developing lip-reading and oral communication skills.

The communication barrier between sign and non-sign language speakers may be reduced in the coming years thanks to the recent advances in neural machine translation and computer vision. Recent works [5,6,9] are making steps towards sign language translation by automatically generating detailed human pose skeletons from spoken language. Skeletons are represented by 2D/3D coordinates of human joints also known as *keypoints*; given a set of estimated keypoints, one can visualize them as a wired skeleton connecting the modeled joints (see the middle row of Figure 1). Although such visualizations are theoretically useful for understanding sign language, no studies have been made so far on whether they are indeed understood by deaf people.

In this work, we study *if and how well members of the Deaf community understand automatically generated sign language videos*. Apart from skeleton visualizations, we go one step further and generate realistic videos using the state of the art human motion transfer method Everybody Dance Now (EDN) [2]. We run a study with four native sign language speakers and record their understanding of both skeleton visualizations and generated signing videos by asking three different types of feedback: a global classification of the video in terms of topic, a translation into American English, and a final subjective rating about how understandable the videos were. We further quantitatively study the quality of the videos generated using the EDN method via the percentage of keypoints one can re-estimate on the generated frames.

For signing videos and keypoints, we utilize a subset of the recent How2Sign [3] dataset, a large dataset of American Sign Language (ASL) signing videos. Our

---

main results indicate that a) the generated videos were generally preferred over the skeleton visualizations and that b) the current state of the art in image generation is not good enough for sign language translation out-of-the-box. Specifically, we show that the model struggles with generating the hands, which play a central role in sign language understanding.

2   Methodology and Results

Generating realistic signing videos. To generate an animated video of a signer given a set of keypoints, we use the Everybody Dance Now (EDN) [2] approach. It is worth noting that this approach models facial landmarks separately, something highly desirable in our case as they are one of the critical features for sign language understanding. The input and output is shown in Figure 1. We use OpenPose [1] to extract keypoints (middle row) from the source video (top row); the keypoints are then used to condition a Generative Adversarial Network (GAN) that generates each video frame, using the model from [7].

Our model was trained on a subset of the How2Sign dataset [3] that contains videos from *two* professional American Sign Language interpreters. Specifically, keypoints extracted from videos of the first signer (top row in Figure 1) were used to learn the model that generates realistic videos of the second signer (bottom row). The training dataset consists of more than 28 hours of sign language translations from instructional videos.

Quantitative Results. An approximate but automatic way of measuring the visual quality of the generated videos is by measuring the number of keypoints that can be reliably detected by OpenPose in the source and generated videos. We focus only on the 125 upper body keypoints which are visible in the How2Sign videos and discard those from the legs. We use two metrics: a) the Percentage of Detected Keypoints (PDK), which corresponds to the fraction of keypoints from the source frame which were detected in the synthesized frame and b) the Percentage of Correct Keypoints (PCK) [8], which labels each detected keypoint as "correct" if the distance to the keypoint in the original image is less than 20% of the torso diameter.

In Table 1 we present these metrics for different OpenPose confidence thresholds. We report results for all keypoints, as well as when restricting the evaluation only on the hand keypoints, as this is a very important part of sign language understanding. We see that although in general the repeatability of keypoints is high, when focusing on the hands, the model fails to generate reliable keypoints, something that may severely hinder sign language understanding.

How understandable are the generated signing videos? We evaluate the degree of understanding for both skeleton visualizations and generated videos by showing 3-minute-long videos to four native ASL speakers. Two watched the generated videos, while the other two watched the animated 2D skeletons visualizations. During the evaluation, each subject was asked to: a) classify six videos between ten categories of instructional videos; b) answer the question *"How well could you understand the video?"* on the five-level scale (Bad, Poor, Fair, Good,

Table 1. Percentage of Detected Keypoints (PDK) and Percentage of Correct Keypoints (PCK) for all keypoints (125 for upper body, hands and face) and just for the hands (21 for each hand), when thresholding at different detection confidence scores.

	PDK			PCK		
min. detection confidence	0	0.2	0.5	0	0.2	0.5
All keypoints	0.99	0.88	0.87	0.90	0.94	0.96
Hands	0.99	0.38	0.17	0.18	0.23	0.26

Fig. 1. Sample of the source video (top row) used to automatically extract 2D keypoints with OpenPose [1] (middle row) and generate frames for a target identity (bottom row). A sample of the generated video can be seen at: `https://youtu.be/inqCQ2Zdh8g`.
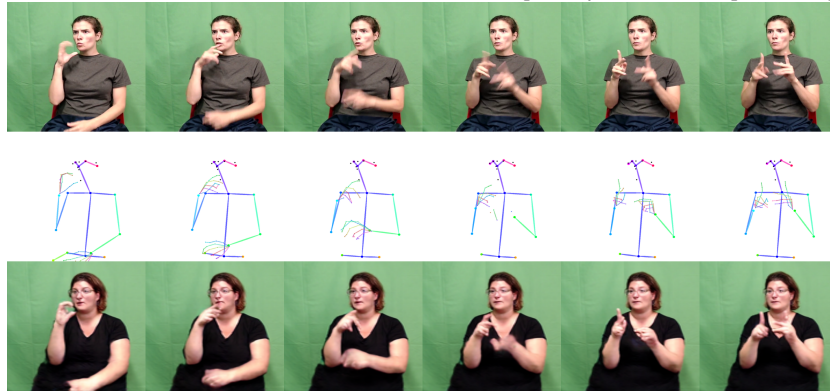
Excellent); c) watch two trimmed clips from the previously watched video, and translate them into American English. Results averaged over all subjects are presented in Table 2. We report accuracy for the classification task, the Mean Opinion Score (MOS) for the five-scale question answers and BLEU [4] scores for the American English translations. Qualitative results are shown in Table 3.

Overall, results show a preference towards the generated videos rather than the skeleton ones as the former result to higher scores across all metrics. In terms of general understanding of the topic, it seems that both visualizations did relatively well; we see that subjects were able to mostly classify the videos correctly. When it comes to finer grained understanding, however, as measured via the English translations, we see from both Tables 2 and 3 that the translation task cannot be solved neither with the skeletons nor with the generated video.

3   Conclusions

In this paper we investigate how well members of the Deaf community actually understand keypoint-based visualizations, the output of choice for many recent automatic sign language translation works. Through our study, we show that subjects prefer synthesized realistic videos over skeleton visualizations; we also show that out-of-the-box synthesis methods are not really effective enough and that subjects struggled to understand the signing videos. We partially attribute

Table 2. Comparison between skeletons and generated videos in terms of classification (accuracy), mean opinion score (MOS) and translation (BLEU) [4].

	Accuracy	MOS	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Skeleton visualization	83.3 %	2.50	10.90	3.02	1.87	1.25
Generated video	**91.6 %**	**2.58**	**12.38**	**6.71**	**3.32**	**1.89**

Table 3. Groundtruth and translations for two clips from a "Food and Drinks" class. All subjects were able to correctly identify the class. Two subjects watched the skeleton visualizations, while two different subjects watched the videos generated by EDN.

Groundtruth	**I'm not going to use a lot, I'm going to use very very little.**
Skeleton	That is not too much don't use much, use a little bit
EDN	Don't use a lot, use a little dont use lot use little bit
Groundtruth	**I'm going to dice a little bit of peppers here.**
Skeleton	cooking chop yellow peppers
EDN	cook with a little pepper chop it little bit and sprinkle

References

1. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE TPAMI (2019)
2. Chan, C., Ginosar, S., Zhou, T., Efros, A.: Everybody dance now. In: ICCV (2019)
3. Duarte, A.C.: Cross-modal neural sign language translation. In: ACM-MM (2019)
4. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: ACL (2002)
5. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: ECCV (2020)
6. Stoll, S., Camgoz, N.C., Hadfield, S., Bowden, R.: Text2sign: towards sign language production using neural machine translation and generative adversarial networks. In: IJCV (2020)
7. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)
8. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE TPAMI **35**, 2878–90 (12 2013)
9. Zelinka, J., Kanis, J.: Neural sign language synthesis: Words are our glosses. In: WACV (2020)