How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language

Amanda Duarte[1,2⋆], Shruti Palaskar[3], Deepti Ghadiyaram[5], Kenneth DeHaan[4], Florian Metze[3,5], Jordi Torres[1,2], and Xavier Giro-i-Nieto[1,2]

[1]Universitat Politècnica de Catalunya, [2]Barcelona Supercomputing Center, [3]Carnegie Mellon University, [4]Gallaudet University, [5]Facebook AI

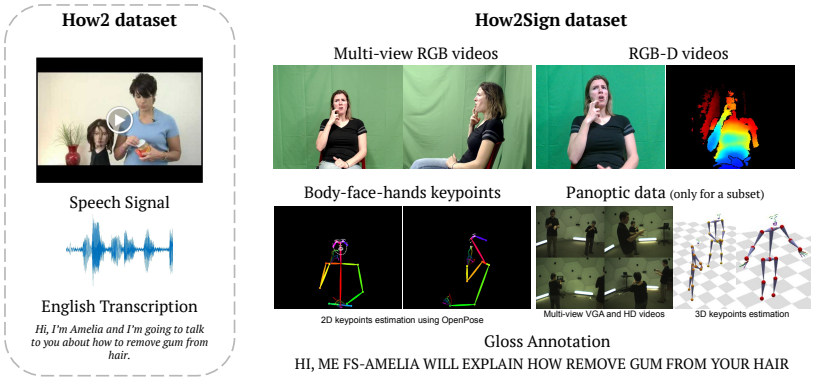

Fig. 1. Samples of data included in the **How2Sign** dataset.

1 Introduction

Sign Languages are the primary means of communication for an estimated 466 million deaf or hard-of-hearing people worldwide [1]. One of the most important factors that has hindered the progress in the areas of automatic sign language recognition, generation, and translation is the absence of large annotated datasets [2], especially *continuous sign language* datasets, *i.e.* datasets that are annotated and segmented at the sentence or utterance level.

An important factor for the lack of datasets is that collection and annotation of continuous sign language data is a laborious and expensive task. It needs to be done by linguistic experts together with a native speaker, e.g a Deaf person. Although there are datasets available from linguistic sources [14,12,7,6] and sign language interpretations from broadcast [5,17,10], they are usually weakly annotated or have a small vocabulary size, and often lack all the modalities required for cross-modal sign language translation research (e.g spoken language and the corresponding aligned sign language translation). Table 1 provides an overview

⋆ *Corresponding author:* amanda.duarte@upc.edu
Presented as an extended abstract at the Sign Language Recognition, Translation & Production workshop (SLRTP) at European Conference on Computer Vision 2020.

of publicly available datasets for continuous sign language in comparison with the How2Sign, our work-in-progress dataset collection.

How2Sign is a large-scale collection of multi-view and multimodal signing videos in American Sign Language (ASL) for over 2500 instructional videos from the existing How2 dataset [13]. Figure 1 shows sample of the data contained in the dataset. Working in close collaboration with native ASL speakers and professional interpreters, we collected approximately **80 hours** of multi-view (multiple RGB and a depth sensor) signing videos and corresponding gloss annotations[1] [11]. In addition, a three-hour subset was further recorded in a geodesic dome setup using hundreds of cameras and sensors, which enables detailed 3D reconstruction and pose estimation and paves the way for vision systems to understand the 3D geometry of sign language.

Table 1. Publicly available, continuous sign language datasets. SL refers to the sign language used, *trans.* refers to translation of the signing videos in its correspondent spoken language, *gloss* to gloss annotations [11] and *speech* to a parallel speech track.

Name	SL	Vocab.	Duration(h)	Content		
				trans.	gloss	speech
Video-Based CSL [7]	Chinese	178	100	✓		
SIGNUM [15]	German	450	55	✓	✓	
RWTH-Phoenix-2014T [4]	German	3k	11	✓	✓	
DGS-Korpus [8]	German	–	50	✓	✓	
Boston104 [16]	ASL	104	8.7 (min)	✓	✓	
How2Sign (ours)	ASL	16k	79	✓	✓	✓

2 The How2Sign dataset

The How2Sign dataset consists of a parallel corpus of instructional videos and their corresponding American Sign Language translation (ASL) videos and annotations. 80 hours of multi-view ASL videos were collected, as well as gloss [11] annotations. The multiple data sources allow for high-quality automatic annotations of 2D body, face and hand keypoints that will be made available together with the videos. For the subset that was recorded in the Panoptic studio [9], accurate 3D keypoints will also be made available.

Table 2. How2Sign dataset statistics. Keypoints for green screen studio were estimated by OpenPose [3]. The number of unique signers is 11.

	Green screen studio			Panoptic studio	
	train	val	test	val	test
Videos	2,213	132	184	48	76
Duration (h)	69.62	3.91	5.59	1.14	1.82
Sentences	31,128	1,741	2,322		
Vocabulary size	15,686	3,218	3,670		
Signers not in train set		0	1	2	2

The instructional videos that were translated to ASL come from the existing *How2 dataset* [13], a publicly available large-scale multi-modal dataset that covers a variety of topics with utterance-level time alignments between the speech and the ground-truth English transcription. We selected a 60-hour subset of the How2 300h set for the How2Sign training set, and used the complete 300h-subset validation and test sets as the How2Sign validation and test sets, respectively.

[1] work in progress.

Detailed statistics are presented in Table 2. The collected corpus contains video recordings by 11 different signers, covering *more than 35k sentences*, with a English vocabulary of more than 16k different English words.

As shown in Table 2, the validation and test sets explicitly contain videos from two signers that are not present in the training set. We envision this to be used for measuring the *generalization across different signers*. Moreover, a subset (approx. 70 min) of the test set has multiple ASL translation videos for the same instructional video, recorded both with a signer that appears in the training set and with a signer that does not.

Recording Setup. To collect American Sign Language translation videos we collaborated with 11 signers. The following sample group self-identified as: 45% hearing (n=5), 36% Deaf (n=4), and 18% hard-of-hearing (n=2).

The video were recorded in a supervised setting, in two different studios: the green screen studio and the Panoptic studio, both presented below. We recorded the complete 80 hours of the dataset in the *green screen studio* setting. We then further collected duplicate recordings in the muti-view studio for a smaller subset of videos from the validation and test splits (approx. 3h in total).

The *green screen studio* was equipped with a depth and a high definition (HD) camera placed in frontal view of a green screen, and another HD camera placed at a lateral view. All three cameras recorded videos at 1280x720 resolution, at 30 frames per second. The *Panoptic Studio* [9] is a system equipped with 480 VGA cameras, 31 HD cameras and 10 RGB-D sensors all synchronized. All cameras were mounted over the surface of a geodesic dome [2], providing redundancy for weak perceptual processes (such as pose detection and tracking) and robustness to occlusion. In addition to the multiview VGA and HD videos, the recording system also estimated 3D skeletons poses of the interpreters, that will also be made publicly available.

Recording pipeline. Before recording the ASL translations for each video, the signer would watch the video and read the transcript as subtitles. After that, they were asked to performed the translation into ASL while watching the video with subtitles at a slightly slower-than-normal (0.75) speed. For each hour of video recorded, preparation, recording and video review required a 3 hour process on average. The dataset was recorded in 65 days within a period of 6 months.

Privacy, Bias and Ethical Considerations: Privacy. All research steps followed procedures approved by an Institutional Review Board including a Human Subjects Research training done by the researchers and a consent form provided by the participants agreeing on being recorded and making their data available for research purposes.

Data distribution and bias. In order to create the data as balanced as possible as well as a signer independent dataset, we distribute the signers in the recordings across the different splits.

Geographic. All the participants were born and raised in the USA and learned ASL as their primary or second language at school time.

[2] http://www.cs.cmu.edu/~hanbyulj/panoptic-studio/

Signer variety. Our dataset was recorded with the collaboration of 11 signers with different body proportions. Six of them were self-identified male and five self-identified female.
Data bias. Our data does not contain large diversity in race/ethnicity, skin tone, background scenery, lighting conditions and camera quality.

Acknowledgements

This work received funding from Facebook through gifts to Carnegie Mellon University and Universitat Politècnica de Catalunya. Amanda Duarte has received support from la Caixa Foundation (ID 100010434) under the fellowship code LCF/BQ/IN18/11660029. Shruti Palaskar was supported by the Facebook Fellowship program. This work has also received funding by the project TEC2016-75976-R of the Spanish Ministerio de Economa y Competitividad and the European Regional Development Fund. The authors would like to thank Chinmay Héjmadi, Xabier Garcia and Brandon Taylor for their help during the data collection and processing.

References

1. 2019, W.H.O.: Deafness and hearing loss, https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss 1
2. Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al.: Sign language recognition, generation, and translation: An interdisciplinary perspective. In: ACM SIGACCESS Conference on Computers and Accessibility. pp. 16–31 (2019) 1
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018) 2
4. Cihan Camgoz, N., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: CVPR. pp. 7784–7793 (2018) 2
5. Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: CVPR. pp. 2568–2574. IEEE (2009) 1
6. Dreuw, P., Forster, J., Deselaers, T., Ney, H.: Efficient approximations to model-based joint tracking and recognition of continuous sign language. In: International Conference on Automatic Face & Gesture Recognition. pp. 1–6. IEEE (2008) 1
7. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: AAAI (2018) 1, 2
8. Jahn, E., Konrad, R., Langer, G., Wagner, S., Hanke, T.: Publishing dgs corpus data: Different formats for different needs. In: LREC. pp. 83–90 (1981) 2
9. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3334–3342 (2015) 2, 3
10. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015) 1

11. Liddell, S.K., et al.: Grammar, gesture, and meaning in American Sign Language. Cambridge University Press (2003) 2
12. Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, A.L., Braffort, A., Glauert, J., Safar, E.: Dicta-sign–building a multilingual sign language corpus. In: LREC. Istanbul (2012) 1
13. Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., Metze, F.: How2: a large-scale dataset for multimodal language understanding. arXiv preprint arXiv:1811.00347 (2018) 2
14. Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., Cormier, K.: Building the british sign language corpus. Language Documentation & Conservation pp. 136–154 (2013) 1
15. Von Agris, U., Kraiss, K.F.: Signum database: Video corpus for signer-independent continuous sign language recognition. In: Workshop on Representation and Processing of Sign Languages. pp. 243–246 (2010) 2
16. Zahedi, M., Dreuw, P., Rybach, D., Deselaers, T., Ney, H.: Continuous sign language recognition-approaches from speech recognition and available data resources. In: Workshop on Representation and Processing of Sign Languages (2006) 2
17. Zelinka, J., Kanis, J., Salajka, P.: Nn-based czech sign language synthesis. In: International Conf. on Speech and Computer. pp. 559–568. Springer (2019) 1