

Supplementary Material for Sign Stitching: A Novel Approach to Sign Language Production

Harry Walsh
harry.walsh@surrey.ac.uk

Ben Saunders
b.saunders@surrey.ac.uk

Richard Bowden
r.bowden@surrey.ac.uk

CVSSP
University of Surrey
Guildford, UK

First, we provide figures to show the produced sequences from the sign stitching approach. Then we explain how we collect both the isolated and continuous dictionaries. Finally, we provide additional implementation details.

1 Stitching Example

Here we use the ground truth gloss labels and timings to show we can accurately recreate a real continuous sequence. On the BSL Corpus T dataset, we use the SignBank dataset as our dictionary which has gloss variant labels. Allowing us to select the same form of each sign.

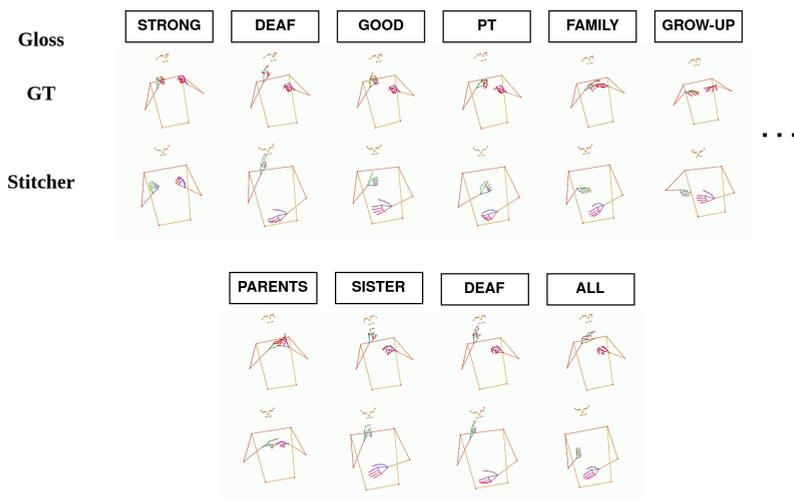


Figure 1: An example of the sign stitching approach on the BSL Corpus T.

2 Translation Examples

Here we share a translation example. Showing each step of the pipeline from Text-to-Gloss (T2G) (rows 1 and 2), then Gloss-to-Pose (T2P) (rows 2 and 6) and finally Pose-to-Sign (P2S) (rows 6 and 7). The approach is able to recreate the sequence, however with a small variation in the sign style due to the different forms of the signs in our dictionary. Here we are limited to only comparing against the Progressive Transformer (PT), as it is the only model which is publicly available.



Figure 2: A Text-to-Sign (T2S) translation example on the RWTH-PHOENIX-Weather-2014T dataset. Showing the original spoken language (Text), the corresponding glosses (Gloss), the original video (GT Video), the ground truth pose (GT pose), the output from the Progressive Transformer (PT), the produced stitched sequence (Stitcher), and the output from the SignGAN module (SignGAN).

Note in columns 2 and 3 the approach is using a different form of the Sign ‘WEST’, hence the different motion but the same handshape.

Video examples can be found here, https://github.com/walsharry/Sign_Stitching_Demos.

3 Dictionary

In the experiments, we tested two different dictionaries: 1) collected from isolated examples, and 2) a dictionary created from continuous data. Next, we provide further details about each:

Isolated: Here, the signs are sourced from individuals who perform each sign in isolation, typically starting from and returning to a resting position. When experimenting on the BSL Corpus T (BSLCPT) we use the Signbank dataset [1], it contains over 3,000 signs

and includes all the lexical variants found in the BSLCPT dataset. However no such dictionary exists for the RWTH-PHOENIX-Weather-2014T (PHOENIX14T) and Meine DGS Annotated (mDGS) dataset, therefore we collect a dictionary from a range of sources such as [8]. We find the mDGS dataset has a target vocabulary of 10,801. However, without the gloss variant, we find the core gloss vocabulary reduces to 4,434. We collect a total of 7,206 signs to experiment with. We use the method described in Section 3.2 (Stitching), step 1 to overcome issues with an incomplete vocabulary. To create word embeddings we apply Facebook’s implementation of Fasttext [8]. When experimenting on the BSLCPT we use the English implementation whereas on PHOENIX14T and mDGS we use the German implementation.

Continuous: Here we create a dictionary using the gloss timing annotations. The signs are taken from the test and dev data only, so that the back translation model has not seen the signs during training. As the examples come from the continuous sign, we omit the cropping step of the stitching pipeline. These dictionaries have an abundance of signs to choose from when stitching. We filter the dictionary and remove short signs as these are most likely co-articulated and therefore not suitable or out of context. We opt to randomly select the first sign in the sequence, and subsequent signs are chosen to ensure the most natural transition. Therefore, we select the sign from the dictionary in which the location of the wrist is closest to the last frame of the previous sign.

4 Implementation Details

In our experiments, we conducted a grid search for optimal hyper-parameters and identified the following settings as the most effective. Our encoder-decoder translation model is constructed with an embedding size of 512 and a feed-forward size of 1024. We find that the optimal number of layers and heads is dataset-dependent, with smaller datasets requiring fewer layers compared to mDGS where we use 3 layers and 4 heads. The models utilise dropout with a probability of 0.1 [8], ReLU activations between layers [8], and pre-layer normalisation for regularisation and training stability. Training employs a ‘reduce on plateau’ scheduler with a patience of 5 and a decrease factor of 0.8. The layers are initialised using a Xavier initializer [8] with zero bias, and during training, Adam optimization is employed [8]. The initial learning rate is set to 10^{-4} , and we train the model until convergence. During decoding, we utilise a greedy search algorithm. The loss scaling factors, λ_y , λ_d , λ_f and λ_C are set to 1.0, 0.1, 0.3 and 0.2, respectively. When stitching we enforce a minimum sign length, U_{min} of 4 frames. The cropping threshold, α_{crop} is dataset-dependent, we find values between the range of 0.1 to 0.35 most effective. All sequences are subsampled to 12 frames per second (fps) for computational efficiency.

For each dataset, we create a dictionary of 500 facial expressions. We scale the counter loss by 0.01 and we set an embedding dimension of 512. The encoder and decoder are initialized with the same settings as our translation model.

The angular pose representation comprises 104 angles, while the Euclidean representation consists of 61 keypoints (21 for each hand and 19 for the body and face). The face mesh we add includes 128 keypoints, which are a subset of Mediapipe’s 478-face mesh [8].

For comparison, we train a progressive transformer on each dataset until convergence using the parameters from [8].

4.1 Datasets

The approach is tested on three datasets, the Public Corpus of German Sign Language, 3rd release, the mDGS dataset [9], PHOENIX14T [10] and the BSLCPT [11]. BSLCPT contains 211 participants from 8 regions in the UK, performing 4792 individual signs from a range of age groups. The participants perform narrative, interviews and participate in free conversation. Similarly, mDGS contains 330 participants engaging in free-form signing. Whereas, the PHOENIX14T dataset is extracted from German TV weather broadcasters and contains over 8,000 parallel sequences.

4.2 Duration Generation

On the mDGS and BSLCPT we use the gloss time stamp annotations to generate the target duration’s for training. However, when ground truth timing information is not available, such as on the PHOENIX14T dataset, we propose a novel sign segmentation approach based on the stitching method described in Section 3.2 (Stitching). Given the ground truth gloss labels, we generate the stitched sequence, P_{stitch} , but without step 4 (sign resampling).

Comparing the stitch sequence and the ground truth, we find that the motion can vary due to different lexical variants present compared to our dictionary. However, we find that the handshape is often still consistent. So, we take the keypoints that correspond to the signer’s hands and normalise the rotation so that the index finger metacarpal bone is fixed on the y-axis and the palm is fixed on the xy-plane, giving P_{stitch}^H, P^H . Our next step is to align the two sequences so that we can infer the duration of the signs in the ground truth. For this we apply Dynamic Time Warping (DTW), such that;

$$A_{i,j} = DTW(P_{stitch}^H, P^H) \quad (1)$$

As we know the duration of the isolated signs in the stitched sequence, by analysing the alignment path, A_j , we can infer the duration of the signs in the original ground truth sequence.

To evaluate this segmentation approach we calculate the duration for each gloss in the mDGS dataset test set. We achieve a sign level frame F1-score of 0.6373 a similar score compared to [9] that achieves a top score of 0.63. Validating that our stitching approach can also be used for sign segmentation. It is worth noting our approach requires gloss information, but is computationally inexpensive compared to the LSTM used in [9].

References

- [1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.
- [2] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018.
- [3] Kearsy Cormier, Jordan Fenlon, Trevor Johnston, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam, and Bencie Woll. From corpus to lexical database to online dictionary: Issues in annotation of the bsl corpus and the development of bsl signbank. In *5th Workshop on the Representation of Sign Languages: Interactions*

- between Corpus and Lexicon [workshop part of 8th International Conference on Language Resources and Evaluation, Turkey, Istanbul LREC 2012. Paris: ELRA. pp. 7–12, 2012.*
- [4] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.
- [5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- [6] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release, 2020. URL <https://doi.org/10.25592/dgs.corpus-3.0>.
- [7] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.
- [8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013.
- [9] Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. Linguistically motivated sign language segmentation. *arXiv preprint arXiv:2310.13960*, 2023.
- [10] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, 2020.
- [11] Adam Schembri. British sign language corpus project: Open access archives and the observer’s paradox. In *sign-lang@ LREC 2008*, pages 165–169. European Language Resources Association (ELRA), 2008.
- [12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 2014.