WLASL-LEX: a Dataset for Recognising Phonological Properties in American Sign Language

Federico Tavella and **Viktor Schlegel** and **Marta Romeo**
Aphrodite Galata and **Angelo Cangelosi**
{name.surname}@manchester.ac.uk
Department of Computer Science, The University of Manchester

Abstract

Signed Language Processing (SLP) concerns the automated processing of signed languages, the main means of communication of Deaf and hearing impaired individuals. SLP features many different tasks, ranging from sign recognition to translation and production of signed speech, but it has been overlooked by the NLP community thus far. In this paper, we bring attention to the task of modelling the phonology of sign languages. We leverage existing resources to construct a large-scale dataset of American Sign Language signs annotated with six different phonological properties. We then conduct an extensive empirical study to investigate whether data-driven end-to-end and feature-based approaches can be optimised to automatically recognise these properties. We find that, despite the inherent challenges of the task, graph-based neural networks that operate over skeleton features extracted from raw videos are able to succeed at the task to a varying degree. Most importantly, we show that this performance pertains even on signs unobserved during training.

1 Introduction

Around 200 languages in the world are signed rather than spoken, featuring their own vocabulary and grammatical structures. For example the American Sign Language (ASL) is not a mere translation of English into signs and is unrelated to the British Sign Language (BSL). Their non-textual nature introduces many challenges to their automated processing, compared with purely textual NLP. Research on Sign Language Processing (SLP) encompasses tasks such as sign language detection, i.e. recognising if and which signed language is performed (Moryossef et al., 2020) and sign language recognition (SLR) (Koller, 2020), i.e. the identification of signs either in isolation or in continuous speech. Other tasks concern the translation from signed to spoken (or written) (Camgoz et al., 2018) language or the production of signs from text (Rastgoo et al., 2021). With the recent success of deep learning-based approaches in computer vision (CV), as well as advancements in —from the CV perspective—related tasks of action and gesture recognition (Asadi-Aghbolaghi et al., 2017), SLP is gaining more attention in the CV community (Zheng et al., 2017).

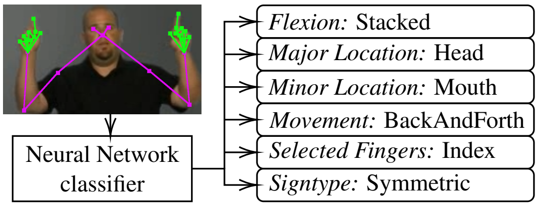

Figure 1: We annotate ASL sign videos with their corresponding phonological information and skeleton features of the speakers, and train neural networks to recognise the former from the latter.

Due to the complexity of the tasks, some recent approaches to various SLP tasks implicitly rely on *phonological* features (Tornay, 2021; Metaxas et al., 2018; Gebre et al., 2013; Tavella et al., 2021). Surprisingly, however, little work has been carried out on explicitly modelling the phonology of signed languages. This presents a timely opportunity to investigate signed languages from the perspective of computational linguistics (Yin et al., 2021). In the context of signed languages, phonology typically distinguishes between manual features, such as usage, position and movement of hands and fingers, and non-manual features, such as facial expressions. Sign language phonology is a matured field with well-developed theoretical frameworks (Liddell and Johnson, 1989; Fenlon et al., 2017; Sandler, 2012). These phonological features, or *phonemes*, are drawn from a fixed inventory of possible configurations which is typically much smaller than the vocabulary of signed languages (Borg and Camilleri, 2020). For example, there is only a limited number of fingers that can be used

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 453 - 463
May 22-27, 2022 ©2022 Association for Computational Linguistics

to perform a sign due to anatomical constraints. Hence, different signs share phonological properties and well performing classifiers can be used to predict those properties for signs unseen during training. This potentially holds even across different languages, because, while different languages may dictate different combinations of phonemes, there are also significant overlaps (Tornay et al., 2020).

Finally, these phonological properties have a strong discriminatory power when determining signs. For example, in ASL-Lex (Caselli et al., 2017), a lexicon which also captures phonology information, the authors report that more than 50% of its 994 described signs have a unique combination of only six phonological properties and more than 80% of the signs share their combination with at most two other signs. By relying on this phonological information from resources such as ASL-Lex, many signs can be uniquely determined. This means that well performing classifiers can leverage this information to predict signs without having encountered them during training. This is a capability that current data-driven approaches to SLR lack by design (Koller, 2020). Thus, in combination, mature approaches to phonology recognition can facilitate the development of sign language resources, for example by providing first-pass silver annotations for new sign languages based on their phonological properties. This is an important task for both documenting low-resource sign languages as well as rapid developing of large-scale datasets, and for fully harnessing data-driven CV approaches.

To spur research in this direction, we extend the preliminary work by Tavella et al. (2021) and introduce the task of Phonological Property Recognition (PPR). More specifically, with this paper, we contribute *(i)* WLASLLex2001, a large-scale, automatically constructed PPR dataset, *(ii)* an analysis of the dataset quality, and *(iii)* an empirical study of the performance of different deep-learning based baselines thereon.

2 Methodology

We address PPR as a classification problem based on features extracted from videos of people speaking SL. Although manual annotation approaches are widely adopted, these are time consuming and require expert knowledge. Instead, we rely on automated dataset construction. On a high level, we cross-reference a large-scale ASL SLR dataset with an ASL Lexicon and annotate videos of signs with their corresponding phonological properties. We then extract skeletal features, by taking advantage of pre-trained deep models from the computer vision community (Rong et al., 2021; Wang et al., 2019). Finally, we train several deep models to classify them as phonological classes.

2.1 Dataset construction

As previously mentioned, ASL-Lex (Caselli et al., 2017) contains phonological features of American Sign Language, such as where the sign is executed, the movement performed by the hand and the number of hands and fingers involved. The latter properties were coded by 3 ASL-versed people. In our work, we are interested in recognising phonological properties from videos of people speaking ASL. Consequently, we aim to construct a dataset, suitable for supervised learning, containing videos labelled with six phonological properties. Specifically, we choose the manual properties with the strongest discriminatory power to determine signs based on their configuration (Caselli et al., 2017):

(i) *flexion:* aperture of the selected fingers of the dominant hand at sign onset,

(ii) *major location:* general location of the dominant hand at sign onset,

(iii) *minor location:* specific location of the dominant hand at sign onset,

(iv) *movement:* the first movement path of the sign,

(v) *selected fingers:* fingers that are moving or are foregrounded during that movement, and

(vi) *sign type:* symmetry of the hands according to Battison (1978).

A detailed description of all the properties is provided in the appendix.

One of the limitations of ASL-Lex is the small number of examples and lack of variety: its first iteration (ASL-Lex 1.0) contains less than 1000 videos, all signed by the same person. While sufficient for educational purposes, these videos are of limited suitability for developing robust classifiers that can capture the diversity of ASL speakers (Yin et al., 2021). To this end, we source videos from WLASL (Li et al., 2020) (Word Level-ASL), one of the largest available SL datasets, featuring more than 2000 glosses demonstrated by over 100 people, for a total of more than 20000 videos. Each sign is performed by at least 3 different signers, which implies greater variability compared to having one gloss performed by only one user. By cross referencing ASL-Lex and WLASL2000 based on corresponding glosses, we can increase the number of samples available to train our models.

Finally, to leverage state of the art SLR architectures that operate over structured input, we enrich each raw video with its extracted keypoints that represent the joints of the speaker. To do so, we use two pretrained models, FrankMocap (Rong et al., 2021) and HRNet (Wang et al., 2019). While these tracking algorithms follow different paradigms, the former extracting 3D coordinates based on a predicted human model and the latter predicting keypoints as coordinates from videos directly, they produce similar outputs. An important distinction is that while FrankMocap estimates the 3D keypoints, HRNet outputs 2D keypoints with associated prediction confidence scores. We use these different models to explore whether different tracking algorithms affect the recognition of phonological classes. We select a subset of features of the upper body, namely: nose, eyes, shoulders, elbows, wrists, thumbs and first/last knuckles of the fingers. These manual features were determined to be the most informative while performing sign language recognition (Jiang et al., 2021b).

Our final dataset, WLASL-Lex2001 (WLASL2000 + ASL-Lex 1.0), is composed of 10017 videos corresponding to 800 glosses, 3D skeletons (x, y, z from FrankMocap and x, y and $score$ from HRNet) labelled with their phonological properties. A characteristic of this dataset is that it follows a long tailed distribution. Due to the nature of language, some phonological properties are more common than others, which means that some classes are more represented than others. On the one hand, the training setup for our models should take this factor into account, but on the other hand, the advantage of training over phonological classes instead of glosses is that different glosses can share phonological classes.

2.2 Models

To estimate the complexity of the dataset, we use the majority-class baseline and the Multi-Layer Perceptron (MLP) as basic deep models. We further use Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) as models capable of capturing the temporal component of videos. As state-of-the-art SLP architectures that have been used to perform SLR, we use the I3D 3D Convolutional Neural Network (Carreira and Zisserman, 2017; Li et al., 2020) able to learn from raw videos, and the Spatio-Temporal Graph Convolutional Network (STGCN) (Jiang et al., 2021b) that captures both spatial and temporal components from the extracted keypoints.

2.3 Experimental Setup

For each phonological property we generate dataset splits and train dedicated models separately. While a multi-class multi-label approach could achieve higher scores, by relying on potential interdependencies of different properties, we chose to model the properties in isolation, to disentangle the factors that affect the learnability of each property. From now on, when we mention the *dataset*, we refer to an instance of the WLASL-Lex 2001 dataset, where labels are the values of a single phonological class.

We make this distinction because we produce six different train, validation and test splits (with a $70 : 15 : 15$ ratio) stratifying on the corresponding phonological property (*Phoneme*). By doing so, we make sure that *(a)* all splits contain all possible labels for a classification target (i.e. phonological property) and *(b)* follow the same distribution. Since we source the videos from WLASL, we have multiple videos representing each gloss, therefore, randomly splitting our data will result in the fact that glosses in the test set might appear in the training set as well, signed by a different speaker. Thus, to investigate how well the models can predict properties on unseen glosses, we also produce label-stratified splits on gloss-level (*Gloss*), such that videos of glosses in the validation and test set do not appear in training data and vice versa. Thus, to summarise, experiments in the *Phoneme* setting aim to evaluate the capability to recognise phonological properties of signs that were already encountered in the training data, but are performed by a different speaker in the test set. Conversely, experiments in the *Gloss* setting aim to evaluate the capability to recognise phonological properties of signs completely *unseen during training*.

We use an I3D model that has been pre-trained on Kinetics-400 (Carreira and Zisserman, 2017)

		FLEXION		MAJLOCATION		MINLOCATION		MOVEMENT		FINGERS		SIGNTYPE	
		A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$
Phoneme	Baseline	50.3	11.1	34.4	20.0	33.9	3.1	35.5	16.7	48.2	11.1	39.3	20
	MLP$_H$	44.1 ± 2.5	11.1	70.3 ± 2.3	64.0	51.6 ± 2.5	28.2	34.5 ± 2.4	18.7	59.4 ± 2.5	25.0	73.9 ± 2.2	52.6
	MLP$_F$	50.3 ± 2.5	11.1	57.8 ± 2.5	46.8	34.3 ± 2.4	9.1	34.3 ± 2.4	18.7	43.4 ± 2.5	12.9	67.0 ± 2.4	42.8
	RNN$_H$	49.0 ± 2.5	30.0	75.8 ± 2.2	72.4	64.3 ± 2.4	46.0	35.1 ± 2.4	29.5	71.0 ± 2.3	46.5	78.7 ± 2.1	58.8
	RNN$_F$	50.3 ± 2.5	11.1	64.6 ± 2.4	54.2	30.3 ± 2.3	4.0	35.4 ± 2.4	18.1	46.5 ± 2.5	12.4	70.9 ± 2.3	46.8
	STGCN$_H$	**62.3 ± 2.4**	**45.0**	**83.2 ± 1.9**	**78.6**	**74.5 ± 2.2**	**63.5**	**63.6 ± 2.4**	**58.2**	**73.8 ± 2.2**	**56.0**	**84.5 ± 1.8**	**69.6**
	STGCN$_F$	43.4 ± 2.5	20.8	70.5 ± 2.3	62.1	53.0 ± 2.5	40.0	45.7 ± 2.5	37.8	63.1 ± 2.4	32.8	73.0 ± 2.2	53.1
	3DCNN	46.5 ± 2.5	13.2	64.3 ± 2.4	55.2	42.3 ± 2.5	18.6	32.9 ± 2.4	20.8	47.5 ± 2.5	14.5	69.5 ± 2.3	44.8
Gloss	Baseline	**53.1**	11.1	35.7	20.0	42.0	5.0	35.2	16.7	47.4	12.5	38.3	20.0
	MLP$_H$	44.6 ± 2.5	15.5	68.1 ± 2.3	56.6	47.3 ± 2.5	19.7	28.4 ± 2.2	19.8	56.2 ± 2.5	22.9	75.3 ± 2.2	50.7
	MLP$_F$	52.8 ± 2.5	11.1	56.6 ± 2.5	42.9	38.3 ± 2.4	10.7	37.1 ± 2.4	21.7	39.3 ± 2.5	12.5	68.4 ± 2.4	41.2
	RNN$_H$	39.6 ± 2.5	18.0	72.8 ± 2.2	67.3	49.3 ± 2.5	26.3	32.2 ± 2.3	24.9	60.7 ± 2.5	32.5	75.4 ± 2.2	53.5
	RNN$_F$	53.0 ± 2.5	11.1	64.1 ± 2.4	52.6	44.4 ± 2.4	17.8	36.7 ± 2.4	20.1	27.3 ± 2.3	12.7	72.0 ± 2.3	46.9
	STGCN$_H$	49.1 ± 2.5	**21.6**	**77.3 ± 2.1**	**70.0**	**55.1 ± 2.4**	**32.7**	**52.5 ± 2.5**	**46.5**	**65.7 ± 2.4**	**34.4**	**76.6 ± 2.1**	**54.4**
	STGCN$_F$	39.0 ± 2.5	14.4	66.7 ± 2.3	60.1	45.1 ± 2.4	21.1	43.1 ± 2.5	34.9	60.0 ± 2.5	29.2	71.3 ± 2.3	47.5
	3DCNN	46.0 ± 2.5	12.8	64.9 ± 2.4	52.0	10.8 ± 1.5	13.6	32.0 ± 2.3	19.3	45.9 ± 2.5	14.7	71.6 ± 2.3	46.3

Table 1: Accuracy (A.) and per-class averaged accuracy ($\overline{A}$) of various models on the test sets of the six tasks. For accuracy, we report the error margin as a confidence interval at $\alpha = 0.05$ using asymptotic normal approximation. We omit error margins for balanced accuracy as the low number of classes results in a small sample size. Additional performance measures are reported in the appendix.

and fine-tune it on raw videos from our datasets. The other models are trained from scratch using keypoints as input. We fix the length of all input to 150 frames, longer sequences are truncated while shorter sequences are looped to reach the fixed length. We select the best performing model based on performance on the validation set and for the final test set performance we train the models on both train and validation sets. For more details on model selection, consult the appendix. We measure both accuracy, to investigate how well models perform in general, and class-balanced accuracy to take into account how well they are able to model different classes of the phonological properties.

3 Results and discussion

The upper half of Table 1 presents the results for the six dataset splits for the *Phoneme* setting, where glosses in test data could have appeared in training data as well. The poor performance of the simple MLP architecture suggests that the tasks are in fact challenging and do not exhibit easily exploitable regularities. Due to its simplicity, it is barely able to reach the baseline for some properties (34% vs. 35% and 44% vs. 50% for *movement* and *flexion* respectively). In particular, MLP classifying based on FrankMocap (MLP$_F$) output is often the worst performing combination. Conversely, STGCN using HRNet output (STGCN$_H$) outperforms other models on all six tasks. In some cases, for example when predicting *movement* or *flexion*, it is the only model which significantly surpasses the majority class baseline. This superior performance is expected, as this specific combination of the STGCN operating over HRNet-extracted keypoints has been shown to be the largest contributor to the SLR performance on the WLASL2000 dataset (Jiang et al., 2021a).

Models that operate over structured input often outperform the 3D CNN, demonstrating the utility of additional information provided by the skeleton features. The results also suggest that models using the HRNet skeleton output outperform those who use FrankMocap, possibly due to the confidence scores produced by HRNet and associated with the coordinates. This difference in performance suggests to conduct a more rigorous study to investigate the impact of different feature extraction methods as a possible future research direction.

The lower half of Table 1 shows the performance of models to predict the phonological properties of unseen glosses (*Gloss*). The performance of all tasks and all models deteriorates, suggesting that their success is partly derived from exploiting the similarities between glosses that appear in training and test data. However, the best model, STGCN$_H$, performs comparably to the *Phoneme*-split, with a drop of less than 10 accuracy points for five of the six tasks.

Often, crowd sourced (Polonio et al., 2018) or automatically constructed datasets such as ours, have a performance ceiling, possibly due to incorrectly assigned ground truth labels or low quality of input data (Chen et al., 2016; Schlegel et al., 2020). To investigate the former, we measure the agreement on videos that all models misclassify using Fleiss' κ. Intuitively, if models consistently agree on a label different than the ground truth, the ground truth label might be wrong. We find that averaged across the six tasks, the agreement is negligible: 0.09 ± 0.06 and 0.11 ± 0.09 for *Phoneme* and *Gloss* split, respectively.

Similarly, for the latter, if all models consistently fail to assign any correct label for a given video (e.g. all models err on a video appearing in the test sets of *movement* and *flexion*), this can hint at low quality of the input, making it impossible to predict anything correctly. We find that this is not the case with WLASL-LEX2001, as videos appearing in test sets of different tasks tend to have a low mutual misclassification rate: 1% and 0.7% of videos appearing in test sets of two and three tasks were misclassified by all models for all associated tasks for the *Phoneme* split. For the *Gloss* split the numbers are 3 and 0% for two and three tasks, respectively. Together, these observations suggest that the models presented in this paper are unlikely to reach the performance ceiling on WLASL-Lex2001 and more advanced approaches could obtain even higher accuracy scores.

4 Conclusion

In this paper, we discuss the task of Phonological Property Recognition (PPR). We automatically construct a dataset for the task featuring six phonological properties and analyse it extensively. We find that there is potential for improvement over our presented data-driven baseline approaches. Researchers pursuing this direction can focus on developing better-performing models, for example by relying on jointly learning all properties, as labels for different properties can be mutually dependent.

Another possible avenue is to investigate the feasibility of using PRR to perform *tokenisation* of continuous sign language speech, by decomposing it into multiple phonemes, which is identified as one of the big challenges of SLP (Yin et al., 2021).

Acknowledgements

The authors would like to acknowledge the use of the Computational Shared Facility at The University of Manchester. The work was partially supported by the UKRI TAS Node on Trust, the US Air Force project THRIVE++ and the H2020 projects TRAINCREASE, eLADDA and PERSEO.

References

Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, Victor Ponce-Lopez, Xavier Baro, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. 2017. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences. *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, pages 476–483.

Robbin Battison. 1978. Lexical borrowing in american sign language.

Mark Borg and Kenneth P. Camilleri. 2020. Phonologically-Meaningful Subunits for Deep Learning-Based Sign Language Recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12536 LNCS:199–217.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2017. Asl-lex: A lexical database of american sign language. *Behavior Research Methods*, 49(2):784–801.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 4, pages 2358–2367, Stroudsburg, PA, USA. Association for Computational Linguistics.

J Fenlon, Kearsy A Cormier, and Diane Brentari. 2017. Sign language phonology. In *Routledge Handbook of Phonological Theory*.

Binyam Gebrekidan Gebre, Peter Wittenburg, and Tom Heskes. 2013. Automatic sign language identification. *2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings*, pages 2626–2630.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021a. Sign Language Recognition via Skeleton-Aware Multi-Model Ensemble.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021b. Skeleton Aware Multimodal Sign Language Recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 3408–3418.

Oscar Koller. 2020. Quantitative Survey of the State of the Art in Sign Language Recognition.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Scott K. Liddell and Robert E. Johnson. 1989. American Sign Language: The Phonological Base. *Sign Language Studies*, 1064(1):195–277.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Dimitris Metaxas, Mark Dilsizian, and Carol Neidle. 2018. Scalable ASL sign recognition using model-based machine learning and linguistically annotated corpora.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-Time Sign Language Detection Using Human Pose Estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12536 LNCS:237–248.

Davide Polonio, Federico Tavella, Marco Zanella, and Armir Bujari. 2018. Ghio-ca: An android application for automatic image classification. In *Smart Objects and Technologies for Social Good*, pages 248–257, Cham. Springer International Publishing.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign Language Recognition: A Deep Survey. *Expert Systems with Applications*, 164:113794.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2021. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*.

Wendy Sandler. 2012. The Phonological Organization of Sign Languages. *Language and Linguistics Compass*, 6(3):162–182.

Viktor Schlegel, Marco Valentino, André Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A Framework for Evaluation of Machine Reading Comprehension Gold Standards. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Federico Tavella, Aphrodite Galata, and Angelo Cangelosi. 2021. Phonology recognition in american sign language.

Sandrine Tornay. 2021. *Explainable Phonology-based Approach for Sign Language Recognition and Assessment*. Ph.D. thesis, Lausanne, EPFL.

Sandrine Tornay, Marzieh Razavi, and Mathew Magimai.-Doss. 2020. Towards Multilingual Sign Language Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6304–6308. IEEE.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. 2019. Deep high-resolution representation learning for visual recognition. *TPAMI*.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. pages 7347–7360.

Lihong Zheng, Bin Liang, and Ailian Jiang. 2017. Recent Advances of Deep Learning for Sign Language Recognition. *DICTA 2017 - 2017 International Conference on Digital Image Computing: Techniques and Applications*, 2017-Decem:1–7.

A Hyperparameters optimization

Table 2 contains all the hyperparameters explored during our experiment over each different model. The best model is the one that maximises the Matthew's correlation coefficient

$$MCC = \frac{TP{\cdot}TN - FP{\cdot}FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

with TP, TN, FP, FN being true/false positive/negative. For the STGCN we use hyperparameters chosen by Jiang et al. (2021a), because initial experiments on our data showed a difference of at most 2% accuracy, which is within the uncertainty estimate. To find the optimal hyperparameters for the other models, we perform Bayesian optimisation over a pre-defined set. We maximise Matthews correlation coefficient (MCC) (Matthews, 1975) on the validation sets of all six tasks. We choose MCC as it provides a good trade-off between overall and class-level accuracy which is necessary due to the unbalance inherently present in our dataset.

Model	Parameters
MLP	number of layers hidden dimension dropout learning rate scheduler step size gamma
RNN	number of RNN layers RNN hidden dimension RNN dropout
STGCN	learning rate number of groups block size, window size scheduler step size dropout warmup epochs
3D CNN	dropout learning rate gamma scheduler step size window size

Table 2: Set of explored hyperparameters for each different model

B Seed dependency

Table 3 illustrates the performance on the test set for each model with respect to chance as measured by training 5 models from different random seeds. The performance difference is negligible suggesting that model training is largely stable with regard to chance.

Model	Accuracy
MLP	74.39 ± 0.35
RNN	79.12 ± 0.46
STGCN	84.12 ± 0.29
3D CNN	69.23 ± 0.93

Table 3: Mean and standard deviation of accuracy of all architectures trained with the HRNet output, measured on the SignType test set and averaged over 5 different random seeds. Results for the 3D CNN are obtained from the validation set.

C Phonological classes description

Tables 4 to 9 describe in detail the meaning of values for all the phonological classes according to ASL-Lex (Caselli et al., 2017).

The cardinality is calculated on WLASL-Lex, which is why some classes that are in ASL-Lex are not represented (i.e., cardinality equal to 0).

D Additional results

Table 10 illustrates additional results for several different metrics. In particular, we report micro- and macro precision/recall and Matthews correlation coefficient. These metrics help to give a better understanding of the classification results, as they are affected more by data imbalance when compared to accuracy.

Value	Definition	Cardinality
imrp	index, middle, ring, pinky finger	4824
imr	index, middle, ring finger	95
mrp	middle, ring, pinky finger	28
im	index, middle finger	1296
ip	index, pinky finger	51
mr	middle, ring finger	0
mp	middle, pinky finger	0
rp	ring, pinky finger	0
i	index finger	2547
m	middle finger	259
r	ring finger	0
p	pinky	407
thumb	thumb	510

Table 4: Values and relative definitions for selected fingers

Value	Definition	Cardinality
Head	Sign is produced on or near the head	3137
Arm	Sign is produced on or near the arm	219
Body	Sign is produced on or near the trunk	1019
Hand	Sign is produced on or near the non-dominant hand	2194
Neutral	Sign is not produced in another location on the body	3448
Other	Sign is produced in another unspecified location on the body	0

Table 5: Values and relative definitions for major location

Value	Definition	Cardinality
1	Fully open: no joints of selected fingers are flexed	5037
2	Bent (closed): non-base joints are flexed	693
3	Flat-open: base joints flexed less than 90 degrees	909
4	Flat-closed: base joints flexed equal to or more that 90 degrees	507
5	Curved open: base and non-base joints flexed without contact	1130
6	Curved closed: base and non-base joints flexed with contact	642
7	Fully closed: base and non-base joints fully flexed	795
Stacked	Stacked: Flexion of selected fingers differs	123
Crossed	Crossed	181

Table 6: Values and relative definitions for flexion

Value	**Definition**	**Cardinality**
HeadTop	Sign is produced on top of the head	20
Forehead	Sign is produced at the forehead	246
Eye	Sign is produced near the eye	616
CheekNose	Sign is produced on the cheek or nose	511
UpperLip	Sign is produced on the upper lip	53
Mouth	Sign is produced on the mouth	431
Chin	Sign is produced on the chin	717
UnderChin	Sign is produced under the chin	74
UpperArm	Sign is produced on the upper arm	39
ElbowFront	Sign is produced in the crook of the elbow	0
ElbowBack	Sign is produced on the outside of the elbow	13
ForearmBack	Sign is produced on the outside of the forearm	32
ForearmFront	Sign is produced on the inside of the forearm	10
ForearmUlnar	Sign is produced on the ulnar side of the forearm	56
WristBack	Sign is produced on the back of the wriset	23
WristFront	Sign is produced on the front of the wrist	0
Neck	Sign is produced on the neck	68
Shoulder	Sign is produced on the shoulder	101
Clavicle	Sign is produced on the clavicle	419
TorsoTop	Sign is produced in the upper third of the torso	0
TorsoMid	Sign is produced in the middle third of the torso	0
TorsoBottom	Sign is produced in the bottom third of the torso	19
Waist	Sign is produced at the waist	34
Hips	Sign is produced on the hips	59
Palm	Sign is produced on the plam of the non-dominant hand	925
FingerFront	Sign is produced on the front of the fingers of the non-dominant hand	99
PalmBack	Sign is produced on the back of the palm of the non-dominant hand	218
FingerBack	Sign is produced on the back of the fingers of the non-dominant hand	186
FingerRadial	Sign is produced on the radial side of the non-dominant hand	410
FingerUlnar	Sign is produced on the ulnar side of the non-dominant hand	40
FingerTip	Sign is produced on the tip of the fingers of the non-dominant hand	158
Heel	Sign is produced on the heel of the non-dominant hand	88
Other	Sign is produced in an unspecified location on the body	707
Neutral	Sign is not produced on or near the body	3390

Table 7: Values and relative definitions for minor location

Value	Definition	Cardinality
One Handed	Sign only recruits one hand	3939
Symmetrical Or Alternating	Sign recruits both hands Phonological specifications for both hands are identical Movement of both hands is either symmetrical or alternating	3358
Asymmetrical Same Handshape	Sign recruits both hands Only the dominant hand moves The location and orientation of the hands may differ, but the other specifications of handshape are the same Non-Dominant hand must be an unmarked handshape (B A S 1 C O 5)	938
Asymmetrical Different Handshape	Sign recruits both hands Only the dominant hand moves The location and orientation of the hands may differ, and the other specifications of handshape are not the same Non-Dominant hand must be an unmarked handshape (B A S 1 C O 5)	1639
Other	Sign violates Battison's Symmetry and Dominance Conditions	143

Table 8: Values and relative definitions for sign type

Value	Definition	Cardinality
Straight	Straight movement of the dominant hand through xyz space	1938
Curved	Single arc movement of the dominant hand through xyz space Hands may or may not make contact with multiple locations	1255
BackAndForth	Sequence of more than one straight or curved movements	3549
Circular	Circular movement of the dominant hand through space Rotation alone does not constitute a circular movement	1129
None	Entire sign (or first free morpheme) does not have a path movement	1748
Other	Sign has another unspecified path movement	398

Table 9: Values and relative definitions for movement

		Flexion					**MajLocation**					**MinLocation**					**Movement**					**Fingers**					**Signtype**				
		P_μ	P_M	R_μ	R_M	MCC	P_μ	P_M	R_μ	R_M	MCC	P_μ	P_M	R_μ	R_M	MCC	P_μ	P_M	R_μ	R_M	MCC	P_μ	P_M	R_μ	R_M	MCC	P_μ	P_M	R_μ	R_M	MCC
Phoneme	Baseline	50.3	5.59	50.3	11.11	0.0	34.4	6.88	34.4	20.0	0.0	33.87	1.06	33.87	3.12	0.0	35.46	5.91	35.46	16.67	0.0	48.17	5.35	48.17	11.11	0.0	39.32	7.86	39.32	20.0	0.0
	MLP$_H$	44.1	24.5	44.1	20.7	14.6	70.3	65.8	70.3	64.0	58.9	51.6	37.3	51.6	28.2	41.6	34.5	28.0	34.5	26.9	13.9	59.5	29.6	59.5	25.0	37.7	73.9	54.1	73.9	52.6	62.5
	MLP$_F$	50.3	5.6	50.3	11.1	0.9	57.8	52.3	57.8	46.8	41.2	34.3	13.9	34.3	9.1	17.9	34.3	13.1	34.3	18.7	5.7	43.4	17.5	43.4	12.9	4.6	67.1	38.1	67.1	42.8	52.7
	RNN$_H$	49.0	32.1	49.0	30.0	25.4	75.8	75.2	75.8	72.4	66.4	64.3	54.3	64.3	46.0	57.4	35.1	30.1	35.1	29.5	15.9	71.0	53.3	71.0	46.5	56.6	78.7	61.2	78.7	58.8	69.4
	RNN$_F$	50.3	5.6	50.3	11.1	0.0	63.9	56.7	63.9	52.2	50.1	30.3	4.7	30.3	4.0	4.9	35.4	21.4	35.4	18.1	5.2	46.5	9.2	46.5	12.4	8.5	70.9	60.6	70.9	46.8	58.3
	GTN$_H$	62.4	55.4	62.4	45.0	43.9	83.2	80.6	83.2	78.6	76.8	74.5	66.7	74.5	63.5	69.8	63.6	62.1	63.6	58.2	52.7	73.8	71.7	73.8	56.0	61.1	84.5	74.9	84.5	69.6	77.7
	GTN$_F$	43.4	23.6	43.4	20.8	15.3	70.5	66.4	70.5	62.1	58.9	53.0	43.9	53.0	40.0	43.8	45.7	40.8	45.7	37.8	28.6	63.1	39.0	63.1	32.8	44.3	73.0	56.8	73.0	53.1	61.1
	3DCNN	46.5	17.8	46.5	13.2	5.4	64.3	57.2	64.3	55.2	50.3	42.3	22.8	42.3	18.6	29.1	32.9	23.4	32.9	20.8	7.5	47.5	17.8	47.5	14.5	14.6	69.5	44.9	69.5	44.8	55.6
Gloss	Baseline	53.03	5.89	53.03	11.11	0.0	35.69	7.14	35.69	20.0	0.0	42.03	2.1	42.03	5.0	0.0	35.21	5.87	35.21	16.67	0.0	47.38	5.92	47.38	12.5	0.0	38.28	7.66	38.28	20.0	0.0
	MLP$_H$	44.6	18.6	44.6	15.5	8.3	68.1	62.0	68.1	56.6	55.5	47.3	16.8	47.3	13.1	32.5	28.4	20.4	28.4	19.8	4.9	56.2	21.4	56.2	20.3	32.0	75.3	50.6	75.3	50.7	64.3
	MLP$_F$	52.8	5.9	52.8	11.1	−2.1	56.7	46.2	56.7	42.9	39.5	38.3	11.7	38.3	7.9	18.1	37.1	15.9	37.1	21.7	12.5	39.3	10.4	39.3	11.1	0.4	68.4	37.7	68.4	41.2	54.3
	RNN$_H$	39.6	19.8	39.6	18.0	10.9	72.8	68.0	72.8	67.3	62.4	49.3	19.6	49.3	17.5	36.7	32.2	25.7	32.2	24.9	11.3	60.7	36.9	60.7	32.5	40.3	75.4	55.0	75.4	53.5	64.6
	RNN$_F$	53.0	5.9	53.0	11.1	0.0	64.1	57.3	64.1	52.6	50.5	44.4	15.1	44.4	12.3	27.9	36.7	11.2	36.7	20.1	10.0	27.3	10.6	27.3	12.7	3.0	72.0	41.3	72.0	46.9	60.4
	GTN$_H$	49.1	25.6	49.1	21.6	18.9	77.3	72.1	77.3	70.0	68.6	55.1	25.1	55.1	23.3	43.4	52.5	49.4	52.5	46.5	38.0	65.7	37.2	65.7	30.6	47.8	76.6	54.9	76.6	54.4	66.2
	GTN$_F$	39.0	15.1	39.0	14.4	4.7	66.7	63.2	66.7	60.1	53.9	45.1	15.7	45.1	13.2	31.1	43.1	36.0	43.1	34.9	25.8	60.0	32.5	60.0	29.2	39.4	71.3	47.6	71.3	47.5	58.5
	3DCNN	46.0	12.0	46.0	12.8	4.5	65.0	57.5	65.0	52.0	51.8	10.8	12.0	10.8	9.7	9.5	32.0	18.7	32.0	19.3	6.0	45.9	15.1	45.9	14.7	10.7	71.6	46.3	71.6	46.3	58.7

Table 10: Micro-averaged (μ), macro-averaged (M) precision (P) and recall (R) and Matthews correlation coefficient (MCC) of various models on the test sets of the six tasks. We omit error margins as the low number of classes results in a small sample size.