CISLR: Corpus for Indian Sign Language Recognition

Abhinav Joshi **Ashwani Bhat**
Pradeep S **Priya Gole** **Shreyansh Agarwal** **Shashwat Gupta**
Ashutosh Modi
Indian Institute of Technology Kanpur (IIT-K)
ajoshi@cse.iitk.ac.in, ashubhat44@gmail.com
{spradeep20,priyagole20,shreyansha20,shashwatg20}@iitk.ac.in
ashutoshm@cse.iitk.ac.in

Abstract

Indian Sign Language, though used by a diverse community, still lacks well-annotated resources for developing systems that would enable sign language processing. In recent years researchers have actively worked for sign languages like American Sign Languages, however, Indian Sign language is still far from data-driven tasks like machine translation. To address this gap, in this paper, we introduce a new dataset CISLR (Corpus for Indian Sign Language Recognition) for word-level recognition in Indian Sign Language using videos. The corpus has a large vocabulary of around 4700 words covering different topics and domains. Further, we propose a baseline model for word recognition from sign language videos. To handle the low resource problem in the Indian Sign Language, the proposed model consists of a prototype-based one-shot learner that leverages resource-rich American Sign Language to learn generalized features for improving predictions in Indian Sign Language. Our experiments show that gesture features learned in another sign language can help perform one-shot predictions in CISLR.

1 Introduction

Existing works in natural language processing have shown promising improvements in text classification, translation as well as generation in widely used spoken languages. However, sign language, on the other hand, still lacks sufficient resources for developing models using data-driven approaches, leading to low improvements in tasks like translation and generation. One such low-resource sign language includes Indian Sign Language (ISL). As per the 2011 Indian Census, there are about 6 million deaf people in India (Wikipedia, 2022). According to Ethnologue (a reference publication documenting information about living languages of the world), ISL is the most widely used sign language in the world, and it is 151st most "spoken" language in the world (Ethnologue, 2022).

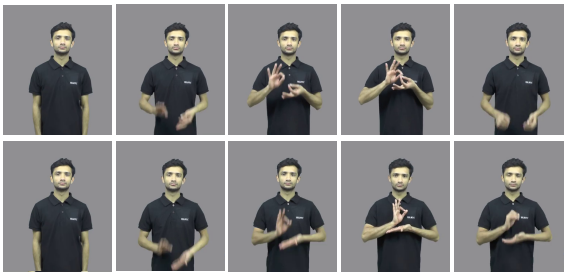

Figure 1: An example of the same signer showing different signs for the same word "Buddhist." Though the movement of the arm is similar, the hand gestures differ.

However, there is a huge deficit of sign language interpreters e.g., according to the Government of India organization Indian Sign Language Research and Training Center (ISLRTC), there are only 300 certified sign language interpreters in India (http://islrtc.nic.in/). This has resulted not only in the widening of the communication barrier between the deaf community and the rest of the population but also has resulted in the very limited development of educational material for sign languages. On the technology side, there are lack of standard benchmarks for ISL resulting in low development and a lack of comparison of machine learning based solutions for ISL, e.g., word recognition, translation, etc. In contrast, relatively speaking, other sign languages (e.g., American Sign Language (ASL), Deutsche Gebärdensprache (DGS)) have a sufficient number of annotated resources for data-driven approaches (Li et al., 2020a; Koller et al., 2015; Sincan and Keles, 2020), and recent multimodal approaches have shown significant improvements in terms of applications (Xu et al., 2022; Albanie et al., 2020; Jiang et al., 2021; Moryossef et al., 2020a).

With an increase in the number of deaf people in India (> 6 million), it is important to develop technologies that could aid in processing ISL and narrow the communication gap between the deaf community and the rest of the population. This

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10357–10366

paper is aimed towards this goal, and via this paper, we draw the attention of the NLP research community to developing technologies for highly low-resource ISL. In this work, we introduce Corpus for Indian Sign Language Recognition (CISLR). The corpus consists of word-level videos of trained ISL signers. The corpus consists of 7050 videos covering 4765 words. Along with the corpus, we introduce the task of word-level Indian Sign Language recognition. The task is challenging due to the limited number of videos for each word and the variability in the signs for the same word by different signers. For example, as shown in Fig. 1, the same signer represents the word "Buddhist" with different hand gestures. Another example is the word "command" where the two signers explain the word; however, they use slightly different gestures (App. Fig. 3). Given that we have limited data for training models for word recognition in ISL, we propose using relatively high-resource languages like ASL (American Sign Language) to transfer knowledge to ISL. Recently, WLASL (Word Level American Sign Language) (Li et al., 2020a), a large-scale word-level sign language dataset, has been released. We try to leverage the WLASL benchmark to extract gesture features. Our experiments show that gesture features learned in another sign language can help perform one-shot predictions in Indian Sign Language. Overall, we make the following contributions in this work:

- We introduce the Corpus for Indian Sign Language Recognition (CISLR). CISLR consists of 4765 words in the form of 7050 videos. Via CISLR, we create a new benchmark for word-level recognition in Indian Sign Language. We release the corpus and model via Github: `https://github.com/Exploration-Lab/CISLR`.
- We perform a detailed analysis of the corpus and provide word-level clustering to understand the diversity and nuances of the dataset.
- We propose a prototype learning method for one-shot prediction in CISLR, leveraging the American Sign Language benchmark to learn rich and generalized gesture features. The results show that it is possible to do knowledge transfer from high-resource sign language to low-resource language.

2 Related Work

There has been active interest in the research community in developing tools and techniques for processing sign languages. Since sign languages contain both visual, gestural, and language modalities, both the vision (Li et al., 2020a) and natural language (Yin et al., 2021) research communities have developed techniques. A number of tasks for sign language processing have been proposed, for example, sign language detection (Moryossef et al., 2020b), identification (Monteiro et al., 2016), segmentation (Bull et al., 2020), recognition (gloss detection) (Imashev et al., 2020; Sincan and Keles, 2020), translation (Moryossef et al., 2021; Yin and Read, 2020a,b; Camgoz et al., 2018, 2020) and generation (Saunders et al., 2020b,a; Xiao et al., 2020). In this paper, we focus on the task of sign language word recognition (gloss detection) from videos. A number of benchmarks have been proposed for gloss detection in sign languages other than ISL (Mesch and Wallin, 2012; Fenlon et al., 2015; Gutierrez-Sigut et al., 2016). A few of the popular sign language resources include (Martinez et al., 2002; Zahedi et al., 2005; Efthimiou and Fotinea, 2007; Gutierrez-Sigut et al., 2016). A large number of benchmarks pose the problem of gloss detection using only RGB images. In contrast, few of the resources (Oszust and Wysocki, 2013; Chai et al., 2015) provide depth modality facilitating the task of gesture recognition using 3D depth maps. In contrast, there are a very few datasets for the Indian Sign Language. Some of the existing datasets include Rekha et al. (2011) which consists of 290 static images for 26 alphabets, Nandy et al. (2010) contains 600 videos corresponding to 22 classes, and Kishore and Kumar (2012) contains 800 videos for 80 different classes. Moreover, the unavailability of these datasets publicly remains a problem for data-driven approaches. INCLUDE dataset (Sridhar et al., 2020) contains 263 classes from 15 different word categories in the form of 4287 videos. Another resource for continuous Indian sign language is ISL-CSLRT (Elakkiya and Natarajan, 2021) which captures 100 sentences in the form of 700 sign videos.

Recently, self-supervised pretraining has attracted attention in the sign language recognition community, Selvaraj et al. (2022) released a large corpus of sign language data for self-supervised pretraining, highlighting the significance of pretraining for both in-language and cross-lingual transfer. Another exciting work in American Sign Language includes WLASL-LEX (Tavella et al., 2022), which explores modeling the phonological

aspects of sign languages. Though the community has started exploring a wide variety of NLP techniques for sign languages, the unavailability of annotated data resources and benchmarks remains the primary challenge for data-driven approaches.

3 CISLR: Corpus for Indian Sign Language Recognition

The major challenge with existing sign language datasets is the small vocabulary size. Limited vocabulary makes it harder to apply deep learning methods for real-world use cases. One of the central motivations for creating a new ISL benchmark is to have a larger vocabulary size to aid further processing. We create CISLR by scraping and curating data from two publicly available internet resources. The first source is the Indian Sign Language Research and Training Center (ISLRTC) (a Government of India initiative: https://islrtc.nic.in/), which provides an ISL Dictionary with the sign language words in the form of videos. Our second source is a non-profit organization IndianSignLanguage (www.indiansignlanguage.org) which offers another collection of Indian Sign Language (ISL) signs in the form of videos. We scrape both publicly available videos from YouTube (www.youtube.com) for both sources. Each video is annotated with an English word.

Data cleaning and Pre-processing: To create a benchmark from the acquired set of videos, we clean the data to remove noise and certain discrepancies. **1) Videos with description:** Since both the sources create a dictionary of words in the ISL, they sometimes contain the description/explanation of the word in sign language along with the word in the video. For our corpus, we only consider the videos containing the words, not the description. We manually check for the videos with the descriptions and remove them from our corpus. **2) Multiple entries for similar words:** In the dictionary, there are videos that are annotated with multiple similar words. For example, the words "restrained", "bound", "confined", and "chained" were mapped to the same sign language video. For our corpus, we remove these multiple entries and create a single class for a video. **3) Additional information in the videos:** Few of the collected dictionary videos contain the signers along with the pictorial representation of the word. For example, in a video containing the sign of the word "apple", a picture of an apple is placed in the video. For our sign language benchmark, we remove the portion containing additional pictures keeping only the signer in a frame. We crop the portions of the frames with the signer showing gestures for the respective word. The example shown in Figure 1 shows the sample of a pre-processed set of frames in CISLR.

Word Categories: CISLR has India-specific words (e.g., names of political parties, organizations, etc) that are not part of standard English. To get a quantitative measure of how much do the words in CISLR overlap with other sign language datasets, we compared the list of words with a large-scale American Sign Language dataset (WLASL) (Li et al., 2020b). We found an overlap of 1282 words with the WLASL dataset. The created dataset exhibits rich diversity in the dictionary words. To qualitatively find the diversity in the dataset, we manually categorize the words into 57 different categories. Out of all the categories, the top 5 categories in terms of the number of words include "action", "geography", "banking", "time", and "flora and fauna". App. Table 5 provides a detailed list of the categories along with the corresponding number of words. The rich number of clusters highlights the diversity of the proposed ISL benchmark, where the average number of words in a cluster is around 123.

Comparison with other Sign-Language datasets: To quantitatively judge the statistics of the created dataset, we compare the curated dataset with the existing sign-language datasets. Overall, CISLR contains the maximum number of words across all the sign language datasets. Table 1 highlights the comparison between various sign language datasets. The more extensive vocabulary makes the curated benchmark more applicable for real-world sign language recognition tasks. The upper part of the table considers only the word-level datasets present in the respective sign languages. In contrast, as Indian Sign Language has a low number of resources, we compare with all the ISL datasets to the best of our knowledge. Note that ISL-CSLRT is a sentence-level dataset. Among the existing datasets for Indian sign language, our dataset has more resources in terms of vocabulary size, and diversity in terms of different signers (71 signers) (also see App. Fig. 6).

Dataset Insights: In the scraped videos for Indian Sign Language, we found that in some of the videos, the signer repeats the sign twice, however,

Datasets	Sign-Language	Words	Videos	Avg. Videos/ Word	Signers	Modalities	Categories
Boston ASLLVD	American	2742	9794	3.6	6	RGB	-
DEVISIGN-L	Chinese	2000	24000	12	8	RGB, depth	-
DGS Kinnect	German	40	3000	75	15	RGB, depth	-
GSL	Greek	20	840	42	6	RGB	-
LAS64	Argentinian	64	3200	50	10	RGB	-
LSE-sign	Spanish	2400	2400	1	2	RGB	-
Perdue RVL-SLLL	American	39	546	14	14	RGB	-
PSL Kinnect 30	Polish	30	300	10	-	RGB, depth	-
RWTH-BOSTON-50	American	50	483	9.7	3	RGB -	
WLASL	American	2000	21,083	10.5	119	RGB	-
Nandy et al. (2010)	Indian	22	600	27.3	-	RGB	-
Kishore and Kumar (2012)	Indian	80	800	10	-	RGB	-
INCLUDE	Indian	263	4287	16.3	7	RGB	15
ISL-CSLRT	Indian	186	700	3.8	7	RGB	-
CISLR (Ours)	Indian	4765	7050	1.5	71	RGB	57

Table 1: The proposed Indian-Sign Language Dataset comparison with other Sign-Language datasets.

there is variation in the repeated sign for the word. Since these videos have the same gloss repeated twice, one can split the videos into two halves and consider both as separate samples of the respective gloss in the dataset. In comparing CISLR with other datasets, we only considered one-half of such videos. To explore further, we created 3 versions of the dataset, "CISLR_v1.0-a", consisting of the first half of such videos, "CISLR_v1.0-b," consisting of the second half; and "CISLR_v1.0-ab" consisting of both halves. Here, we have put v1.0 to mark the current version of CISLR as the first version. In the future, when we will expand the dataset we will update the version number.

Dataset	# Videos	# Prototypes	# Test Samples
CISLR_v1.0-a	7050	4765	2285
CISLR_v1.0-b	7050	4765	2285
CISLR_v1.0-ab	9692	4765	4927

Table 2: Split size of different versions of the created CISLR dataset.

	Top-1	Top-5	Top-10
I3D Classifier	30.2	55.1	63.6

Table 3: The table shows accuracy (%) obtained for classifier trained on WLASL dataset.

4 Task Formulation

The gloss recognition tasks in CISLR can be formally defined as follows. Given a video of a signer (performing gestures and actions), the task is to predict the corresponding gloss label (word). As the proposed CISLR dataset contains a low number of average videos per word (1.5 videos per word), we formulate the gloss recognition task as a one-shot learning task and provide a single sign video of all unique labels in the corpus as prototypes. The remaining samples in the corpus are considered test queries for evaluating gloss recognition. Table 2 provides the split distribution of the proposed CISLR corpus. We consider the standard metric of Top-1, Top-5, and Top-10 classification accuracy scores for evaluation.

5 Baseline Model for Word Recognition

The one-shot recognition task in the proposed CISLR corpus reflects a practical setting where it is not always possible to collect enough videos for a word (creating videos requires experts and is an expensive process), consequently, the task encourages the development of low-resource learning-based techniques for recognition. For baselines, we explore if features obtained from a network trained in high-resource language are useful for recognition in low-resource sign language like ISL. We choose WLASL as it is a large-scale word-level dataset on American Sign Language which provides videos corresponding to 2000 words. Moreover, a large number of signers and a higher number of videos per word make it suitable for training a generalized classifier that can generate features corresponding to different gestures. Table 3 shows the performance of classifiers trained on the WLASL 2000 dataset.

Figure 2 shows the pipeline for representation learning in CISLR. We use the state-of-the-art model on the WLASL dataset Inception3D (I3D) and train it on the 2000 class WLASL dataset.

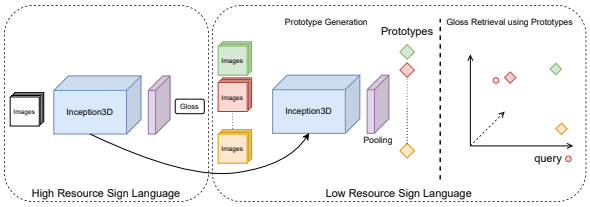

Figure 2: Training in High Resource Sign Language to generate prototypes in a Low Resource Sign Language.

The original I3D network (Carreira and Zisserman, 2017) is trained on ImageNet (Deng et al., 2009) and fine-tuned on Kinetics-400 (Kay et al., 2017). We follow the same strategy as Li et al. (2020b) to fine-tune I3D architecture on the WLASL dataset and obtain the classification scores. We use the penultimate layer of the trained model to generate features corresponding to the prototype videos in the CISLR dataset. For each test sample video, we generate the features using the same I3D network and assign the gloss corresponding to the nearest prototype.

Table 4 shows the results for all three versions of the dataset (CISLR_v1.0-a, CISLR_v1.0-b, and CISLR_v1.0-ab). We observe that a classifier trained in high-resource sign language like WLASL is helpful for sign retrieval in low-resource sign language like CISLR. Moreover, we observe that in the version of CISLR where both the split of the videos were considered (CISLR_v1.0-ab), the retrieval performance is significantly higher than the version with only one split (CISLR_v1.0-a and CISLR_v1.0-b). The reason for such an increase in retrieval performance is the presence of repeated signs by the same signer in the same setting. This improvement in performance highlights that the model trained in WLASL can generalize to slight variations in the sign present in a video. Overall, the Top-1 classification performance of the model trained on WLASL is 30.2% on WLASL's test set, and the same classifier features, when used for the one-shot task in CISLR, perform with 16.81%. This not only shows the use of high-resource sign language for learning low-resource sign language but also highlights the presence of gesture grounding, which could be useful for learning generalized representation in multiple sign languages. An interesting direction to explore in the future would be to do gesture representation learning using data from multiple sign languages. A model that could provide generalized gesture-level features for multiple sign languages would not only help construct advanced NLP models for sign languages but also facilitate the linguistic understanding of sign languages, making them more accessible and reachable to the community.

Dataset	# Test Samples	Top-1	Top-5	Top-10
CISLR_v1.0-a	2285	16.81	20.04	22.58
CISLR_v1.0-b	2285	16.11	19.61	21.97
CISLR_v1.0-ab	4927	43.41	48.06	49.83

Table 4: Performance of prototype features on the created CISLR dataset. (accuracy values are in %)

6 Discussion and Future Directions

Apart from the low resource availability in Indian Sign Language, there are other challenges. Due to the vast diversity of ISL users, we observed the use of slightly different gestures for the same gloss. This makes the gloss recognition task more challenging in a low-resource setting. We also observe dialectical variations in ISL, i.e., the presence of various demography-specific words which are grounded in the region/cultural-specific concepts. Moreover, we also observed that for a few of the words like "author" (explained as a gesture for "book" followed by a gesture for "writing"), the order of gestures is not fixed, and different signers use different order to convey the word "author."

In the future, we plan to expand the corpus by including more words and releasing an updated version of the proposed CISLR. As the proposed dataset is gloss/word level, it limits the application of linguistic analysis of sign language. A possible future work would be to extend the corpus, including sentence-level translations, which would facilitate pretraining on sentence-level translations for better gloss recognition in ISL. Moreover, on the gloss/word level, it would be interesting to explore works like WLASL-LEX (Tavella et al., 2022) for the proposed CISLR corpus.

7 Conclusion

We introduce Corpus for Indian Sign Language Recognition (CISLR) and propose the task of sign word recognition from videos. The dataset is reflective of a practical low-data setting where it is not possible to have multiple videos for a sign. To address this we propose a transfer learning based technique to use high-resource American Sign Language for performing classification on ISL. The results encourage the exploration of one-shot techniques further in this domain.

Limitations

Though the proposed dataset has a large vocabulary size, the problem of low video resources in the dataset remains a significant limitation. The unavailability of resources for modality-hungry natural languages like Sign-Languages poses a real-world problem for the data-driven community. Nevertheless, it encourages the development of techniques that can work in low-data regimes.

Ethics Statement

To the best of our knowledge, our work does not have any ethical considerations. We are performing sign language recognition from videos and people in the videos are of Indian ethnicity. The system is aimed for the Indian population only, so we do not see any biases creeping into the system. The dataset is created from publicly available resources and no copyright is violated.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*.

Hannah Bull, Michèle Gouiffès, and Annelies Braffort. 2020. Automatic segmentation of sign language into subtitle-units. In *European Conference on Computer Vision*, pages 186–198. Springer.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

X Chai, H Wanga, M Zhoub, G Wub, H Lic, and X Chena. 2015. Devisign: dataset and evaluation for 3d sign language recognition. *Technical report, Beijing, Tech. Rep*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Eleni Efthimiou and Stavroula-Evita Fotinea. 2007. Gslc: creation and annotation of a greek sign language corpus for hci. In *International Conference on Universal Access in Human-Computer Interaction*, pages 657–666. Springer.

R Elakkiya and B Natarajan. 2021. Isl-csltr: Indian sign language dataset for continuous sign language translation and recognition. *Mendeley Data*.

Ethnalogue. 2022. The indian sign language.

Jordan B Fenlon, Kearsy Cormier, and Adam C. Schembri. 2015. Building bsl signbank: The lemma dilemma revisited. *International Journal of Lexicography*, 28:169–206.

Eva Gutierrez-Sigut, Brendan Costello, Cristina Baus, and Manuel Carreiras. 2016. Lse-sign: A lexical database for spanish sign language. *Behavior Research Methods*, 48(1):123–137.

Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. 2020. A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The k-rsl. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 631–640.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

PVV Kishore and P Rajesh Kumar. 2012. A video based indian sign language recognition system (inslr) using wavelet transform and fuzzy logic. *International Journal of Engineering and Technology*, 4(5):537.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. 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.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020b. 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.

Aleix M. Martinez, Ronnie B. Wilbur, Robin Shay, and Avinash C. Kak. 2002. Purdue rvl-slll asl database for automatic recognition of american sign language. *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172.

Johanna Mesch and Lars Wallin. 2012. From meaning to signs and back:lexicography and the swedish sign language corpus. In *LREC 2012*.

Caio DD Monteiro, Christy Maria Mathew, Ricardo Gutierrez-Osuna, and Frank Shipman. 2016. Detecting and identifying sign languages through visual features. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 287–290. IEEE.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020a. Real-time sign language detection using human pose estimation. In *European Conference on Computer Vision*, pages 237–248. Springer.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020b. Real-time sign language detection using human pose estimation. In *European Conference on Computer Vision*, pages 237–248. Springer.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.

Anup Nandy, Jay Shankar Prasad, Soumik Mondal, Pavan Chakraborty, and Gora Chand Nandi. 2010. Recognition of isolated indian sign language gesture in real time. In *International conference on business administration and information processing*, pages 102–107. Springer.

Mariusz Oszust and Marian Wysocki. 2013. Polish sign language words recognition with kinect. *2013 6th International Conference on Human System Interactions (HSI)*, pages 219–226.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

J Rekha, J Bhattacharya, and S Majumder. 2011. Shape, texture and local movement hand gesture features for indian sign language recognition. In *3rd international conference on trendz in information sciences & computing (TISC2011)*, pages 30–35. IEEE.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705. Springer.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355.

Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. *INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition*, page 1366–1375. Association for Computing Machinery, New York, NY, USA.

Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. 2022. WLASL-LEX: a dataset for recognising phonological properties in American Sign Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 453–463, Dublin, Ireland. Association for Computational Linguistics.

Wikipedia. 2022. Indo-pakistani sign language — Wikipedia, the free encyclopedia. [Online; accessed 23-June-2022].

Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks*, 125:41–55.

Chenchen Xu, Dongxu Li, Hongdong Li, Hanna Suominen, and Ben Swift. 2022. Automatic gloss dictionary for sign language learners. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 83–92, Dublin, Ireland. Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020a. Attention is all you sign: sign language translation with transformers. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, volume 4.

Kayo Yin and Jesse Read. 2020b. Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989.

Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. 2005. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposium*, volume 3663 of *Lecture Notes in Computer Science*, pages 401–408, Vienna, Austria.

Appendix

A Hyperparameters and Training

We use PyTorch (Paszke et al., 2019) for training and development of our architecture. Our architecture trained on the WLASL dataset has 14,337,264 trainable parameters and takes around 30 minutes to train for 1 epoch on the NVIDIA A40 GPU. We use the adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.00001 for finetuning the I3D model on WLASL dataset.

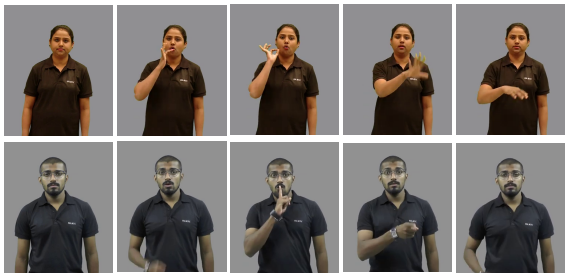

Figure 3: An example of two different signers showing different signs for the same word “command." Though the movement of the arm is similar for both the signers, the hand gestures differ entirely.

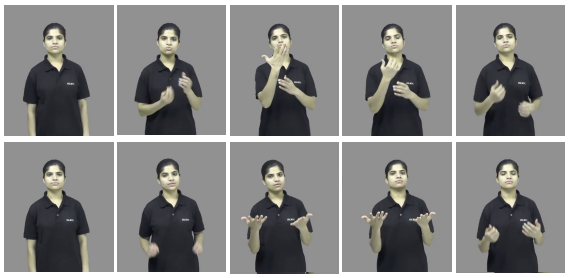

Figure 4: An example of same signer showing different signs for the same word “start.” Though the movement of the arm is similar, the hand gestures differ.

Figure 5: An example of same signer showing different signs for the same word “triumph.” Though the movement of the arm is similar, the hand gestures differ.

Category	# Sign Language Videos
abstract	233
action	718
astronomical	31
attribute	243
authority	54
banking	358
behavior	106
biology	200
chemistry	21
colour	14
commerce	21
comparator	66
computer	42
construction	145
education	47
electronic and electrical	114
emotion	72
entertainment	26
event	117
flora and fauna	300
foods and drinks	227
geography	426
gesture	25
greeting	14
group	65
human	83
instrument	144
internet	56
language	89
legal	184
machine	55
math	132
medical	135
metal and minerals	40
metric	65
miscellaneous object	88
organisation	87
physics	173
politics	23
process	121
profession	206
quality	98
quantity	83
readable	55
relation	113
religious	69
sensation	55
sound	48
spatial position	94
sport	58
stage	48
substance	61
time	304
transport	113
utility item	277
wearable	171
weather	36
weather	36

Table 5: The table shows word video counts of various categories in the created CISLR Dataset.

Figure 6: An image showing the diversity of signers in the acquired corpus for Indian Sign Language.