Continuous Fingerspelling Dataset for Indian Sign Language

Kirandevraj R[1] **Vinod K Kurmi**[2] **Vinay P Namboodiri**[3] **CV Jawahar**[1]
[1] IIIT Hyderabad, India.
[2] IISER Bhopal, India.
[3] University of Bath, UK.
kirandevraj.r@research.iiit.ac.in, vinodkk@iiserb.ac.in,
vpn22@bath.ac.uk, jawahar@iiit.ac.in

Abstract

Fingerspelling enables signers to represent proper nouns and technical terms letter-by-letter using manual alphabets, yet remains severely under-resourced for Indian Sign Language (ISL). We present the first continuous fingerspelling dataset for ISL, extracted from the ISH News YouTube channel, in which fingerspelling is accompanied by synchronized on-screen text cues. The dataset comprises 1,308 segments from 499 videos, totaling 70.85 minutes and 14,814 characters, with aligned video-text pairs capturing authentic coarticulation patterns. We validated the dataset quality through annotation using a proficient ISL interpreter, achieving a 90.67% exact match rate for 150 samples. We further established baseline recognition benchmarks using a ByT5-small encoder-decoder model, which attains 82.91% Character Error Rate after fine-tuning. This resource supports multiple downstream tasks, including fingerspelling transcription, temporal localization, and sign generation. The dataset is available at the following link: https://kirandevraj.github.io/ISL-Fingerspelling/.

1 Introduction

Sign languages serve as the primary communication medium for over 70 million deaf individuals worldwide, yet technological support for these languages remains vastly underrepresented compared to spoken languages. Fingerspelling serves as a critical bridging mechanism in sign languages, allowing signers to spell words from spoken languages letter-by-letter using a dedicated manual alphabet (Padden and Gunsauls, 2003). While American Sign Language (ASL) has benefited from substantial datasets, with recent collections encompassing millions of characters and hundreds of hours of data that enable significant advances in recognition accuracy (Georg et al., 2024), research on ISL fingerspelling recognition has been severely limited by the absence of comparable resources.

The structure of manual alphabets varies across sign languages; some employ one-handed configurations (such as the American Sign Language), while others utilize two-handed systems (such as the ISL). Despite being a subset of the broader sign language lexicon, fingerspelling plays a substantial role in communication. Recognition of fingerspelled sequences presents significant computational challenges owing to two primary factors: first, the movements are characterized by rapid, subtle articulations with extensive co-articulation between consecutive letters, making visual parsing difficult (Patrie and Johnson, 2011); second, fingerspelling predominantly encodes out-of-vocabulary items, including proper nouns, technical terminology, and domain-specific vocabulary, which lack established sign equivalents, limiting the applicability of lexicon-based recognition approaches (Padden and Gunsauls, 2003). This signifies a dedicated focus on fingerspelling.

The development of automated ISL fingerspelling recognition systems has been severely constrained by the absence of large-scale standardized benchmark datasets. Although ASL benefits from substantial resources such as FSboard (Georg et al., 2024) with over 3 million characters and ChicagoFSWild+ (Shi et al., 2019) with 55,232 sequences from 260 signers, existing ISL datasets primarily focus on isolated sign recognition or continuous sentence-level translation tasks(Joshi et al., 2023, 2024), with limited attention to fingerspelling as a distinct recognition challenge.

To address this critical gap in ISL processing resources, we present the first dedicated benchmark dataset for continuous Indian Sign Language fingerspelling recognition, comprising 1,308 fingerspelling segments extracted from 499 ISH News YouTube videos. The dataset totals 70.85 minutes of video data across 1,308 annotated segments containing 14,814 characters, capturing authentic coar-

Proceedings of the Workshop on Sign Language Processing (WSLP), pages 37–42

Figure 1: ISH News fingerspelling example showing eight frames of the word "formaldehyde." The side panel displays letters (F-O-R-M-A-L-D-E) that are sequentially synchronized with the signer's hand configurations. These visual cues serve as our annotation source.

ticulation patterns in naturalistic signing contexts. We establish baseline recognition results using a ByT5-small encoder-decoder transformer model, achieving 82.91% Character Error Rate after fine-tuning and providing reference performance metrics for future research. We have made our dataset and annotations publicly available to facilitate reproducible research and support the broader development of ISL processing technologies for deaf and hard-of-hearing communities.

2 Related Works

Fingerspelling recognition has been extensively studied for American Sign Language using datasets such as ChicagoFSVid (Kim et al., 2016), ChicagoFSWild (Shi et al., 2018), ChicagoFSWild+ (Shi et al., 2019) with 55,232 sequences from 260 signers, and FSboard (Georg et al., 2024) with over 3 million characters. Recent work has extended to fingerspelling span detection in longer videos (Shi et al., 2022; R et al., 2022), enabling automatic localization of fingerspelling segments. In contrast, Indian Sign Language fingerspelling research has primarily focused on image-based handshape classification (Suchithra et al., 2025; Langote et al., 2024), recognizing static handshapes from single frames rather than addressing temporal dynamics in continuous sequences.

Indian Sign Language research has witnessed significant growth with several dataset contributions. Large-scale translation datasets include iSign (Joshi et al., 2024) with 118k video-English pairs and ISLTranslate (Joshi et al., 2023) with 31k pairs from educational videos. Isolated sign recognition is supported by INCLUDE (Sridhar et al., 2020) (263 signs, 4,287 videos), ISL-CSLTR (Elakkiya and Natarajan, 2021) (700 sentence videos, 1,036-word vocabulary), and CISLR (Joshi et al., 2022) (7,050 videos, 4,765 words). However, fingerspelling has been severely underexplored. Existing fingerspelling datasets are exclusively image-based, focusing on isolated alphabet recognition: ISL Fingerspelling (Dongare et al., 2025) provides 14K images, ISL Skeletal (Johnson et al., 2023) contains 3.6K images per letter, ISL Hand Gesture (Biswas, 2024) offers 14.3K images, and Static Gestures of ISL (Singh et al., 2022) include 102K images. None of these captures the temporal dynamics, coarticulation patterns, or continuous sequences necessary for realistic fingerspelling transcription. We address this gap with the first continuous ISL fingerspelling dataset.

3 Fingerspelling Benchmark

3.1 Dataset Creation

We created a continuous fingerspelling dataset from the ISH News YouTube channel by leveraging naturally occurring fingerspelling instances in news videos. The channel employs a distinctive visual cue system where fingerspelling segments, typically proper nouns such as person names and place names, are accompanied by synchronized on-screen text that displays each letter sequentially below a contextual image, timed to match the signer's fingerspelling gestures (Figure 1). We identified 499 unique videos containing these visual cues from which we extracted 1,308 fingerspelling instances. Using the ELAN annotation tool (Brugman and Russel, 2004; Max Planck Institute for Psycholinguistics, 2023), we manually marked the temporal boundaries of each fingerspelling segment around the start and end of the text animation and associated them with the corresponding words or phrases from the visual cues.

Dataset	Type & Size
ISL Fingerspelling (Dongare et al., 2025)	14K images
ISL Skeletal (Johnson et al., 2023)	3.6K img/letter
ISL Hand Gesture (Biswas, 2024)	14.3K images
Static Gestures (Singh et al., 2022)	102K images
Continuous ISL Fingerspelling (Ours)	**1,308 seg. 14,814 chars**

Table 1: Comparison with existing ISL fingerspelling datasets. Prior work focuses on static images of isolated letters, while our dataset provides continuous video sequences.

Validation Outcome	Count
Exact match	136
Signer skipped space	7
Signer made error	5
Too fast to verify	2
Total	**150**

Table 2: Interpreter validation results on 150 randomly sampled segments after correcting validator transcription errors.

This approach enables annotation of continuous fingerspelling sequences from authentic YouTube content.

Video Processing: Following temporal boundary annotation, we preprocessed the video segments to isolate the signer region and remove extraneous visual elements such as the side panel containing text cues. For each annotated segment, we first extracted the corresponding video clips based on marked timestamps. We then employed YOLOv8 (Varghese and M, 2024) person detection on randomly sampled frames to identify the signer's bounding box and select the rightmost detected person (signers consistently appear on the right side of the frame in ISH News videos). To ensure robust cropping across varying camera angles and signer movements, we aggregated bounding boxes across multiple sampled frames using median coordinates. Finally, we applied these computed crop coordinates to extract signer-only video segments, producing 1,308 preprocessed clips containing the signer performing fingerspelling gestures without on-screen text overlays or background elements. This preprocessing ensures that models trained on our dataset focus on visual signing features rather than textual cues.

3.2 Dataset statistics

The dataset comprises 1,308 fingerspelling segments extracted from 499 videos, totaling 70.85 minutes of signing content from 3 unique signers. Among these videos, 408 video IDs overlapped with the iSign (Joshi et al., 2024) sentence-level translation dataset, whereas 91 were not previously included in iSign. The extracted segments contain 14,814 characters total. Alphabets constituted 92.64% (13,724 characters), reflecting the predominantly textual nature of fingerspelling in proper nouns. Spaces accounted for 6.82% (1,011 characters), separating multi-word names and phrases. Numbers appear minimally at 0.17% (25 characters), corresponding to occasional numeric references in names or titles. Other characters comprise 0.36% (54 characters), and primarily include periods used in abbreviations and initials, hyphens in compound names, and occasional parentheses. Table 1 compares our dataset with existing ISL fingerspelling resources, highlighting the shift from static image-based datasets to continuous video sequences.

3.3 Annotation Validation

To validate the reliability of the cue-based annotations, we conducted validation on 150 randomly sampled segments (totaling 8.20 minutes) with a proficient Indian Sign Language interpreter. The interpreter independently transcribed each segment by watching fingerspelling gestures without access to visual cues. In the first round, we identified discrepancies between the cue-based annotations and interpreter transcriptions in 27 cases. Upon closer examination in the second round, we determined that 13 discrepancies resulted from validator transcription errors, which we corrected. The remaining 14 cases reflected actual issues in the source videos or extraction process, as detailed in Table 2. After corrections, 136 of 150 segments (90.67%) achieved exact match with the interpreter validation, confirming the overall reliability of the cue-based annotation approach.

3.4 Fingerspelling Tasks

Our dataset supports three key tasks in sign language processing: **Transcription** converts continuous fingerspelling video segments into character sequences, handling coarticulation, signing speed variations, and ambiguous handshapes. In Section 4, we establish baseline benchmarks for this task. **Temporal Localization** identifies fingerspelling segment boundaries within longer videos. Our annotations provide temporal boundaries for

cue-accompanied fingerspelling instances. The total number of hours of these 499 videos is 20. **Generation** produces signing videos from text with realistic handshapes and transitions. Our dataset can serve as a reference for fingerspelling.

4 Models, Experiments and Results

4.1 Baseline Models

Experimental Setup We conduct two experiments to evaluate fingerspelling recognition performance. First, we evaluated a model pretrained on the iSign dataset (Joshi et al., 2024) in a zero-shot setting on our fingerspelling test set to assess transfer learning from general ISL to fingerspelling. During iSign pre-training, all video IDs overlapping with our fingerspelling dataset were excluded from the training data to prevent data leakage. Second, we fine-tuned the pretrained model on fingerspelling-specific data. We split our fingerspelling dataset based on video ID overlap with iSign: 1,104 segments from videos present in iSign served as the training set, while 204 segments from videos not in iSign formed the test set. The model performance was evaluated using the Character Error Rate (CER).

Model Architecture We adopt the modeling approach from FLEURS-ASL and FSboard (Georg et al., 2024), using a ByT5-small encoder-decoder Transformer. We extracted 75 keypoints (33 body pose, 21 per hand) from MediaPipe Holistic (Lugaresi et al., 2019; Grishchenko and Bazarevsky, 2020) at 15 Hz, yielding 225-dimensional vectors (75 keypoints × 3 coordinates). The iSign dataset provides poses in pose-format (Moryossef et al., 2021). Preprocessing included shoulder-distance normalization for scale invariance, downsampling to 15 Hz, zero-filling for missing keypoints, and padding/truncation to fixed sequence length. We selected ByT5 over subword models because of its character-level tokenization in fingerspelling (Tanzer, 2024). The landmarks were projected through a two-layer feedforward network with layer normalization and dropout into the 1472-dimensional input space of the transformer.

Training We employ a two-stage training strategy: Stage 1 freezes the ByT5 parameters while training only the pose embedding projection for 40 epochs with a learning rate of 1e-4 and batch size of 16, followed by Stage 2 which unfreezes all parameters for end-to-end fine-tuning for 20 epochs with a reduced learning rate of 1e-5 and batch size of 4. We used the AdamW optimizer with gradient clipping, gradient accumulation (steps=2), and 500-step warmup. Training was performed on two RTX 4090 GPUs, completing in approximately 18 hours.

Evaluation Set	**CER (%)**
Pretrained on iSign (zero-shot)	
Test (204 seg.)	432.44
Full dataset (1,308 seg.)	433.06
Fine-tuned on fingerspelling (1,104 train)	
Test (204 seg.)	82.91

Table 3: ByT5-small model performance on fingerspelling transcription. The model was evaluated in zero-shot (pretrained only on iSign) and fine-tuned settings. Test set contains segments from videos not in iSign.

4.1.1 Results

Table 3 presents our baseline results under two evaluation conditions. Without fine-tuning on fingerspelling-specific data, the model pretrained only on general ISL achieved a CER of 432.44% on the test set and 433.06% on the full dataset, demonstrating extremely limited zero-shot transfer capability. After fine-tuning on the fingerspelling training split, performance improved substantially to 82.91% CER, representing an 80.8% relative reduction in error rate. This large performance gap indicates that while learned visual representations from general ISL provide some foundation, fingerspelling recognition requires domain-specific adaptation because of its distinct character-level structure and rapid hand movements. The post-fine-tuning CER of 82.91% establishes a baseline for future work, although it remains substantially higher than the state-of-the-art ASL fingerspelling results (e.g., FSboard achieves 10% CER (Georg et al., 2024)), highlighting the unique challenges and data scarcity for ISL fingerspelling recognition.

5 Conclusion

We present the first continuous fingerspelling dataset for Indian Sign Language, comprising 1,308 video segments from 499 videos totaling 70.85 minutes and 14,814 characters. Our baseline ByT5-small model achieved 82.91% CER after fine-tuning, establishing initial benchmarks while revealing substantial room for improvement. Future work should prioritize expanding dataset scale and signer diversity, investigating transfer learning from larger fingerspelling datasets, and developing improved methods to handle coarticulation patterns in ISL fingerspelling.

Limitations and Ethical Considerations

5.1 Limitations

Our cue-based extraction achieves 90.67% exact match with expert validation after correcting validator errors, with remaining discrepancies from signer errors in videos (5 cases), missing spaces (7 cases), and overly rapid signing (2 cases). The dataset's reliance on ISH News videos with a limited number of professional signers constrains demographic diversity and may reduce generalization to casual or regional signing styles. Temporal boundaries were manually annotated by the first author based on observed correspondence between visual cues and fingerspelling gestures, introducing potential subjectivity in boundary placement. The predominance of proper nouns and news-related terminology may limit model performance in technical jargon or conversational fingerspelling. The relatively small scale (1,308 segments, 70.85 minutes) limits the training of large-scale models and the comprehensive evaluation across diverse fingerspelling scenarios.

5.2 Ethical Considerations

We used publicly available ISH News YoutTube videos, with 407 of 499 videos already in the iSign dataset (Joshi et al., 2024) (which obtained ISH News permission for research use) and the remaining 92 videos featuring identical signers and settings. Sign language videos capture facial expressions and body postures, enabling signer identification and raising privacy concerns despite publicly available nature and institutional permissions. Professional broadcast signers do not represent full ISL community diversity, including regional variations and casual signing styles. Models trained on these broadcast-quality data should not be deployed in accessibility applications without extensive community validation, as generalization gaps could harm deaf users.

Acknowledgements

We are grateful to ISH News for making their sign language news videos publicly available on YouTube. We acknowledge the iSign dataset team for obtaining usage permissions from ISH News, which facilitated our research. We thank the ISL interpreter who assisted with the annotation validation process.

References

Sougatamoy Biswas. 2024. ISL Hand Gesture Dataset.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with elan. In *International Conference on Language Resources and Evaluation*.

Tanvi Dongare, Gaurika Nawani, Aditya Deshpande, Ayaan Shaikh, and Dr. Deepali Javale. 2025. Isl fingerspelling image dataset.

R Elakkiya and B Natarajan. 2021. Isl-csltr: Indian sign language dataset for continuous sign language translation and recognition. *Mendeley Data*, 1.

Manfred Georg, Garrett Tanzer, Saad Hassan, Max Shengelia, Esha Uboweja, Sam S. Sepah, Sean Forbes, and Thad Starner. 2024. Fsboard: Over 3 million characters of asl fingerspelling collected via smartphones. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13897–13906.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic - simultaneous face, hand and pose prediction, on device.

Jans Johnson, Jisha Joseph, Maris Reji, and Megha George. 2023. Indian sign language skeletal-point numpy array using mediapipe.

Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. 2023. Isltranslate: Dataset for translating indian sign language. In *Annual Meeting of the Association for Computational Linguistics*.

Abhinav Joshi, Ashwani Bhat, Pradeep Raj M S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. Cislr: Corpus for indian sign language recognition. In *Conference on Empirical Methods in Natural Language Processing*.

Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. 2024. isign: A benchmark for indian sign language processing. *ArXiv*, abs/2407.05404.

Taehwan Kim, Jonathan Keane, Weiran Wang, Hao Tang, Jason Riggle, Gregory Shakhnarovich, Diane Brentari, and Karen Livescu. 2016. Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation. *Comput. Speech Lang.*, 46:209–232.

Vaishali Langote, Aditya Deshpande, Tanvi Dongare, Gaurika Nawani, Ayaan Shaikh, and Arhaan Mulani. 2024. Bridging the gap: Isl fingerspelling to text, sentiment analysis and language conversion. In *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pages 1–6. IEEE.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *ArXiv*, abs/1906.08172.

Max Planck Institute for Psycholinguistics. 2023. Elan.

Amit Moryossef, Mathias Müller, and Rebecka Fahrni. 2021. pose-format: Library for viewing, augmenting, and handling .pose files. `https://github.com/sign-language-processing/pose`.

Carol Padden and Darline Clark Gunsauls. 2003. How the alphabet came to be used in a sign language. *Sign Language Studies*, 4:10 – 33.

Carol J Patrie and Robert E Johnson. 2011. *RSVP: Fingerspelled word recognition through rapid serial visual presentation*. DawnSignPress.

Prajwal K R, Hannah Bull, Liliane Momeni, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Weakly-supervised fingerspelling recognition in british sign language videos. *ArXiv*, abs/2211.08954.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022. Searching for fingerspelled content in american sign language. In *Annual Meeting of the Association for Computational Linguistics*.

Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2019. Fingerspelling recognition in the wild with iterative visual attention. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5399–5408.

Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Jonathan Michaux, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2018. American sign language fingerspelling recognition in the wild. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 145–152.

Animesh Singh, Sunil K. Singh, Ajay Mittal, and Brij B. Gupta. 2022. Static gestures of Indian Sign Language (ISL) for English Alphabet, Hindi Vowels and Numerals.

Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh M. Khapra. 2020. Include: A large scale dataset for indian sign language recognition. *Proceedings of the 28th ACM International Conference on Multimedia*.

M. Suchithra, Ayushi Gupta, and Abhilasha Kasaraneni. 2025. Fingerspelling for indian sign language using swin transformer. *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 947–952.

Garrett Tanzer. 2024. Fingerspelling within sign language translation. *ArXiv*, abs/2408.07065.

Rejin Varghese and Sambath. M. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6.