Dense Frame Annotations for Low-Resource ISL Fingerspelling Recognition

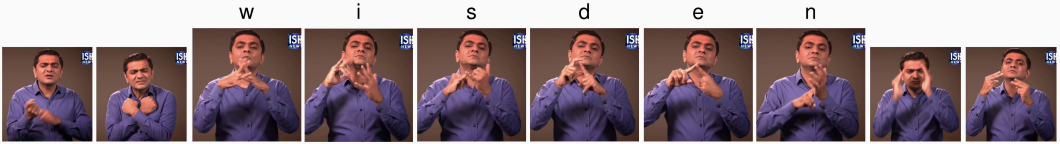
Kirandevraj R[1] Vinod K Kurmi[2] Vinay P Namboodiri[3] C.V. Jawahar[1]

[1] IIIT Hyderabad, India [2] IISER Bhopal, India [3] University of Bath, UK

Presented by **Avijit Dasgupta**, IIIT Hyderabad

Fingerspelling in sign language

People who are crazy for cricket, love to read Wisden which writes about famous cricketers.



Names, technical terms, acronyms

Any word with no dedicated sign

ISL: two-handed alphabet

Limitations of word-level supervision

Existing datasets

`[video]` $\longrightarrow$ `"wisden"`

No letter boundaries.

- Coarticulation blends letters
- Segmentation inferred implicitly
- Breaks down on small data

Total ISL training data: $\sim$71 minutes.

Existing strategies for data scarcity

1. Cross-lingual transfer

Pretrain on ASL or BSL.

Different alphabets, different styles.

2. Weak supervision at scale

Mine subtitles and mouthing.

More labels, same weak signal.

Both add **data**. Neither improves **supervision**.

Our approach: annotation depth over data volume

41 frames

"wisden"

one label / clip

w	i	s	d	e	n

one label / **frame**

Identical training data, **substantially denser supervision**.

Contributions

1. **ISL-FS-Dense**: first character-level frame annotations for any sign language
   125,661 frames · 14,682 characters · **99.1%** expert-validated

2. **Dense supervision wins the comparison**
   4.87 CER vs 16.4 word-level vs 8.1 cross-lingual

3. **Automatic scaling**
   11,700+ discovered segments · 9× the manual set

ISL-FS-Dense

Annotation protocol

Task

- Start and end frame of every character
- 38 unique characters
- 1,308 segments

Challenge: boundary ambiguity

Coarticulation blurs where a letter ends.

Transition-onset rule

Boundary = frame where the hand begins moving to the next letter.

ISL-FS-Dense: statistics and validation

Metric	Value
Segments	1,308
Duration	70.85 minutes
Annotated frames	**125,661**
Character instances	**14,682**
Unique characters	38
Avg. frames per character	8.6
Annotation time	∼35 hours

Interpreter review	Accuracy
Letters in isolation	92.0%
With word context	**99.1%**

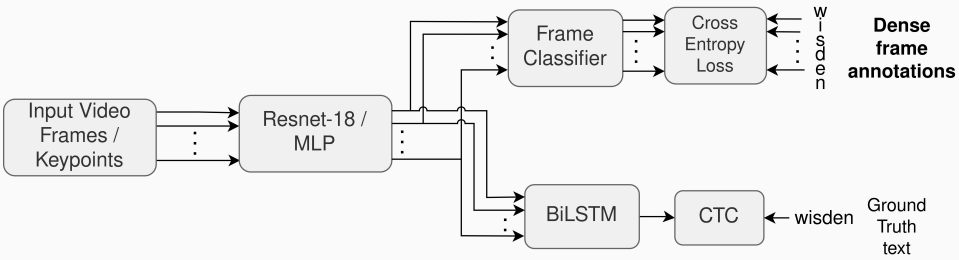
Independent professional ISL interpreter.

660 blind samples

6 genuine annotation errors

Method

Architecture



Frame classifier ← frame labels

CTC branch ← word labels

Shared encoder ⇒ direct ablation.

Two-stage training

Stage 1

Frame classification.

Handshapes first.

Stage 2

Sequence prediction,

frame loss retained.

$$\mathcal{L} = \mathcal{L}_{\text{CTC}} + \lambda \mathcal{L}_{\text{CE}}$$

Controlled comparison

Drop $\mathcal{L}_{\text{CE}} \Rightarrow$ standard word-level supervision. Only the supervision changes.

Experimental setup

Splits (metric: CER)

- Standard (1,104 / 204)
- Signer-independent (810 / 498)
  unseen test signers

Ours

- ResNet-BiLSTM (RGB, 14.3M)
- MLP-BiLSTM (keypoints)

Baselines (all word-level)

- ByT5 (300M)
- Transpeller (BSL fingerspelling)
- Transpeller (BOBSL sign classification, 69M)

Results

Recognition results (CER %, lower is better)

Method	Input	Supervision	CER (%)	
			Standard	Signer-Indep
ByT5 (300M)	RGB	Word	85.9	81.4
Transpeller	RGB	Word	78.7	79.7
Transpeller (BSL pretrained, 69M)	RGB	Word	8.1	**15.7**
ResNet-BiLSTM **(Ours)**	RGB	Word	16.4	37.4
ResNet-BiLSTM **(Ours)**	RGB	Frame+Word	**4.87**	16.8
MLP-BiLSTM **(Ours)**	Keypoint	Word	7.5	21.4
MLP-BiLSTM **(Ours)**	Keypoint	Frame+Word	**5.1**	**14.3**

16.4 → **4.87** (70% rel.) 37.4 → **16.8** (55% rel.)

Analysis

1. Dense supervision yields **large gains** under a controlled comparison
2. **Outperforms cross-lingual transfer** at $\frac{1}{5}$ the parameters
3. BSL *fingerspelling* knowledge does not transfer; *sign classification* does
4. **Keypoints** generalise best to unseen signers

Fingerspelling Detection

Detection without detection supervision

Constraint

Dedicated detectors need non-fingerspelling annotations. We have none.

Observation

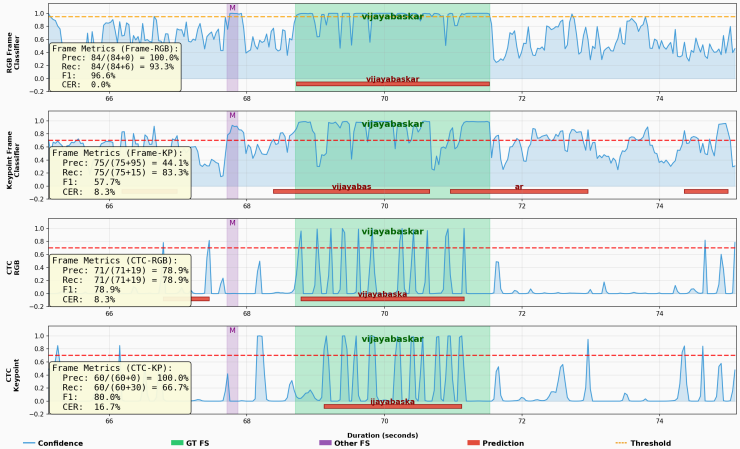
The frame classifier *already* detects.

Confidence = max softmax.

Scores → regions

1. Threshold
2. Merge nearby
3. Drop short

Detection confidence ("vijayabaskar", 10s window)



Frame classifier (top) vs CTC (bottom two).

Detection results

Input	Model	Precision	Recall	F1	CER
RGB	CTC (word-level)	78.0%	71.0%	74.3%	29.2%
	Frame classifier	**83.5%**	**85.7%**	**84.6%**	**11.5%**
Keypoint	CTC (word-level)	70.5%	81.7%	75.7%	27.4%
	Frame classifier	68.0%	**86.4%**	**76.1%**	**17.9%**

84.6% F1, with no detection training.

Automatic Dataset Scaling

Automatic dataset scaling

Source: iSign, 127,105 videos with sentence transcripts only.

Sliding window	Detection-based
Transcribe windows, match transcript words.	Detect regions, match NER proper nouns.
7,833	**7,479**

11,700+ segments, 9× the manual set

Discovered segments improve generalisation

Input	Training Data	Samples	CER (%)
RGB	Manual	810	16.8
RGB	Manual + Discovered	5,464	**15.1**
Keypoint	Manual	810	14.3
Keypoint	Manual + Discovered	5,464	**13.5**

Discovered segments have no frame labels: CTC only.

Weak data helps, once letters are learned densely.

Conclusion

Conclusion

- **ISL-FS-Dense**: first character-level frame annotations, ∼35 hours
- Outperforms word-level (16.4) *and* 10×-larger transfer (8.1) at **4.87** CER
- Detection at **84.6%** F1, with no detection training
- 9× automatic scaling

Thank you

https://kirandevraj.github.io/ISL-Fingerspelling/

kirandevraj.r@research.iiit.ac.in

Backup: training details

With frame supervision

- Pretrain frame classifier, up to 30 epochs
- Joint training, $\lambda = 1$, up to 50 epochs

Without frame supervision

- Freeze encoder, train BiLSTM (30 ep, lr 10^{-4})
- Fine-tune end-to-end (30 ep, lr 10^{-5})

Common

- AdamW, batch size 8
- Gradient clipping (max_norm $= 5.0$)
- Early stopping (patience 10)

Models

- ResNet-18 (ImageNet) $\rightarrow$ 512-dim; signer crop 224×224
- 2-layer BiLSTM, 256 hidden, dropout 0.3 (14.3M)
- Keypoints: 42 points $\rightarrow$ 126-dim $\rightarrow$ MLP $\rightarrow$ 256-dim

Backup: detection hyperparameters

Grid search on a held-out sample of 10 videos.

Method	Threshold	Gap tolerance	Min. duration
RGB frame classifier	0.95	0.3s	0.5s
Keypoint frame classifier	0.70	0.3s	0.5s
CTC RGB	0.70	0.7s	0.5s
CTC keypoint	0.70	0.7s	0.5s

- RGB frame classifier tolerates a much higher threshold; its confidence separates cleanly.
- CTC models need wider gap tolerance to bridge their noisier traces.

Backup: quality of discovered segments

Blind manual transcription, 100 non-overlapping segments per pipeline.

Pipeline	Exact match	Mean CER
Detection-based	79%	7.42%
Sliding window	71%	7.98%

Sources of residual error

1. Signers sometimes fingerspell only *part* of the matched word.
2. Hand occlusions make some letters difficult even for human annotators.

3,867 of 7,479 detection-based segments (51.7%) are not found by sliding window.

Backup: related datasets

ASL

- ChicagoFSWild: 7,304 sequences
- ChicagoFSWild+: 55,232 sequences
- FSboard: 3M+ characters, isolated letters

BSL

- Transpeller: 5,000 manual + ∼100,000 discovered, all word-level

ISL

- Prior work: static handshape images, no temporal dynamics
- ISL-FS: 1,308 continuous segments, word labels

Gap in existing resources

No dataset for *any* sign language provides character-level frame boundaries.