Toward Sign Language Video Understanding in the Real World

Karen Livescu

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

ICASSP 2023

Background

Sign languages

- Meaning expressed through gestures of hands, arms, mouth, eyebrows
- >70 million deaf people, >300 sign languages
- Vocabulary and syntax separate from spoken languages
- No standard written form

https://www.youtube.com/@melmira/featured

Spoken/written language technologies are ubiquitous…

- Automatic speech recognition, translation, search, …

… but not available for sign languages

Technical challenges

- Low-resource, unwritten languages
- Quick motions, coarticulation, inter-signer variability

This talk

I. Sign language background

II. Sign language understanding from video: Data and tasks

III. Toward sign language understanding "in the wild": Case studies from a TTIC/U. Chicago collaboration

Diane Brentari

U. Chicago

Jonathan Keane

U. Chicago

Jonathan Michaux

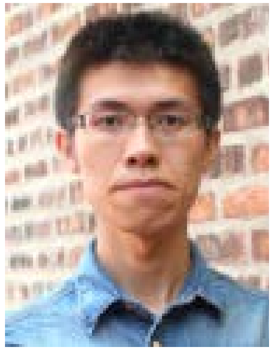
U. Chicago

Aurora Martinez del Rio

U. Chicago

Greg Shakhnarovich

TTIC



Bowen Shi

TTIC

This talk

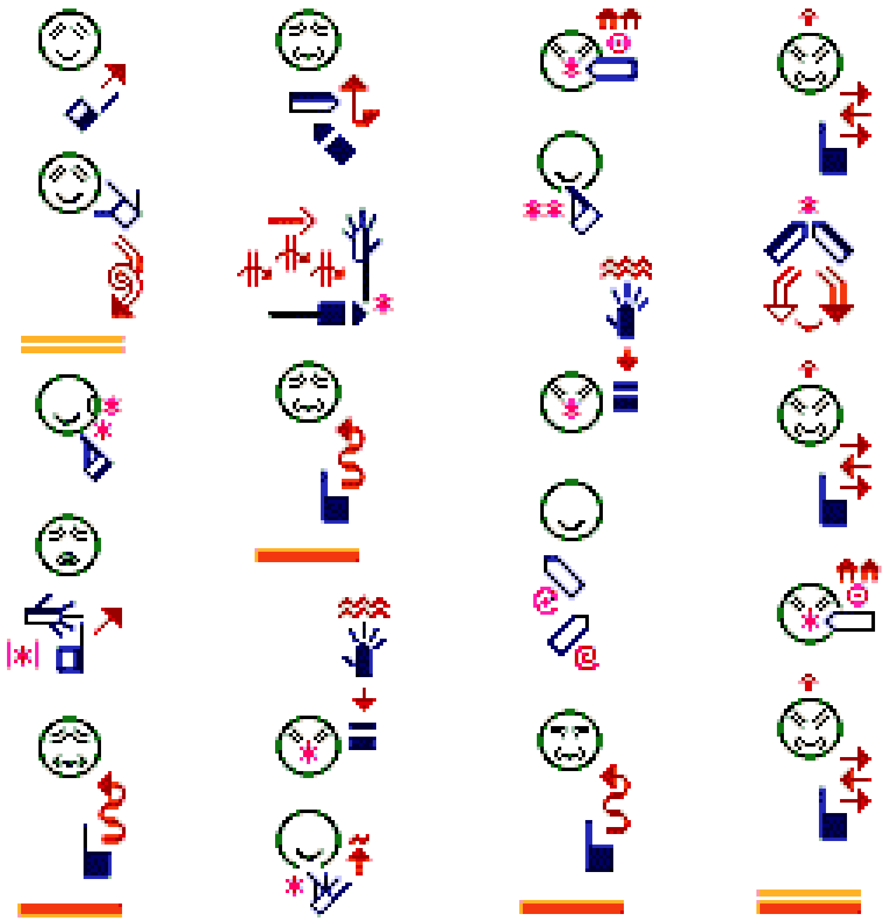
I. Sign language background

II. Sign language understanding from video: Data and tasks

III. Toward sign language understanding “in the wild”: Case studies from a TTIC/U. Chicago collaboration

Background: Sign language transcription systems

- Multiple phonetic, alphabetic, and glossing systems have been developed
- No written transcription system is widely used among signers



Gloss

THERE GOLDILOCKS , HOME ESCAPE
WANDER . ENTER FOREST WANDER .
SMELL FAVORITE SCENT
FOOD SMELL ENJOY WANDER .
WHERE HOUSE WHERE SCENT WHERE ?

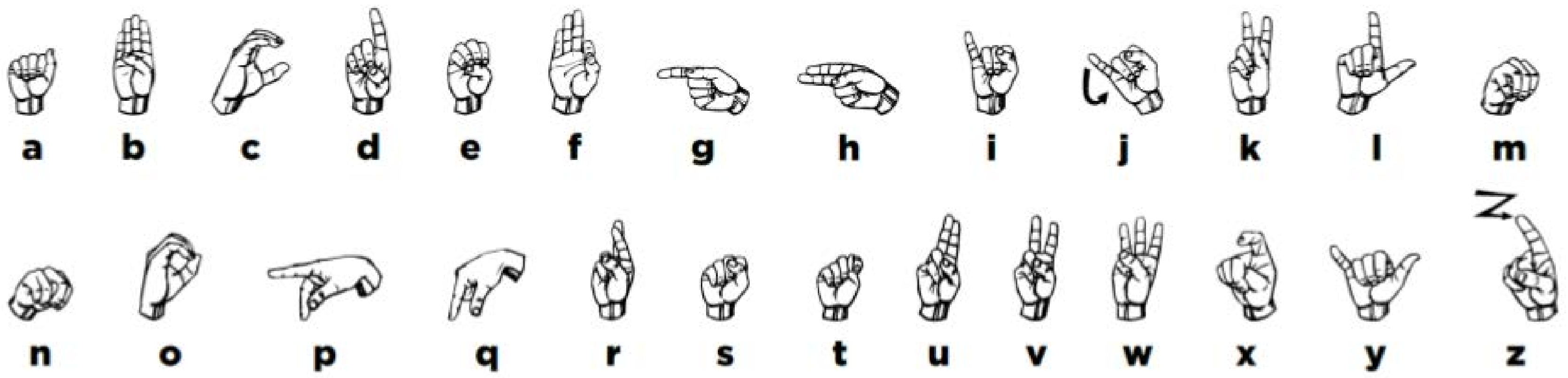
English version

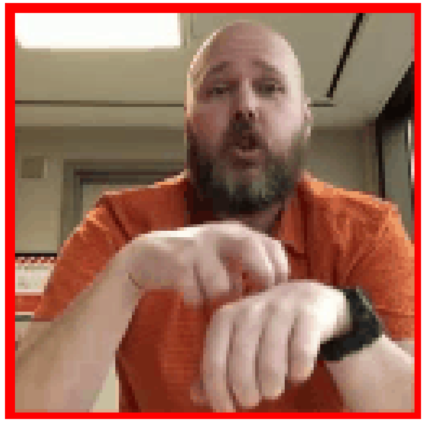
Goldilocks wandered away from her home and into the forest. She smelled the scent of her favorite food and wandered towards the pleasing scent. Where was the house where the scent was coming from?

https://omniglot.com/writing/signwriting.htm

Background: The role of fingerspelling in sign language

- Letter-by-letter signing of a word in a spoken language (e.g., pirate -> P-I-R-A-T-E)
- One handshape/trajectory corresponding to each letter
- Example: Fingerspelling alphabet for American Sign Language (ASL)





(video)

- 12-35% of ASL signs (Padden & Gunsauls 2003)
- Used for important content: names, organizations, emphasized words

➔ For open-domain sign language understanding, crucial to transcribe the fingerspelling

Background: Sign language video media

Interpretation of spoken broadcasts

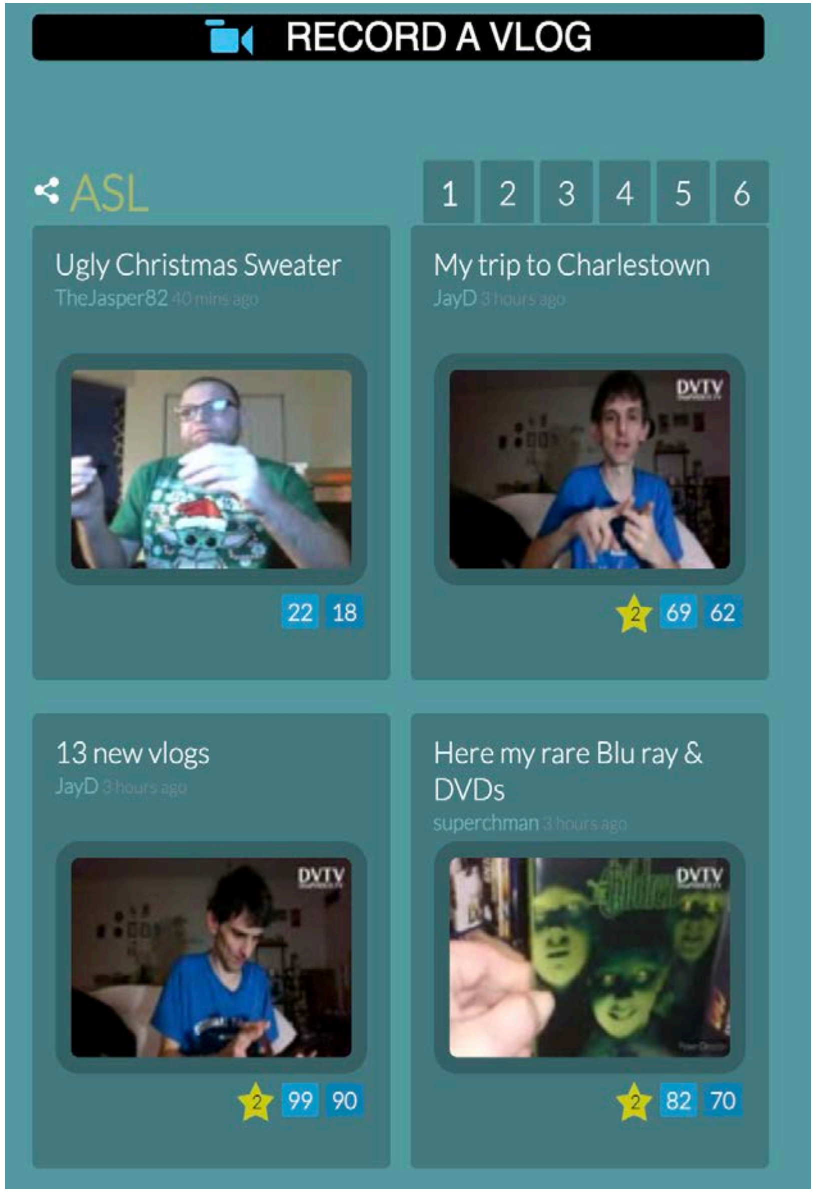
https://washingtonpost.com
Forster et al., “Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather,” LREC 2014

Background: Sign language video media

Native (non-interpreted)
sign language media

News produced in sign language

Vlogs



https//sign1news.com , http://deafvideo.tv

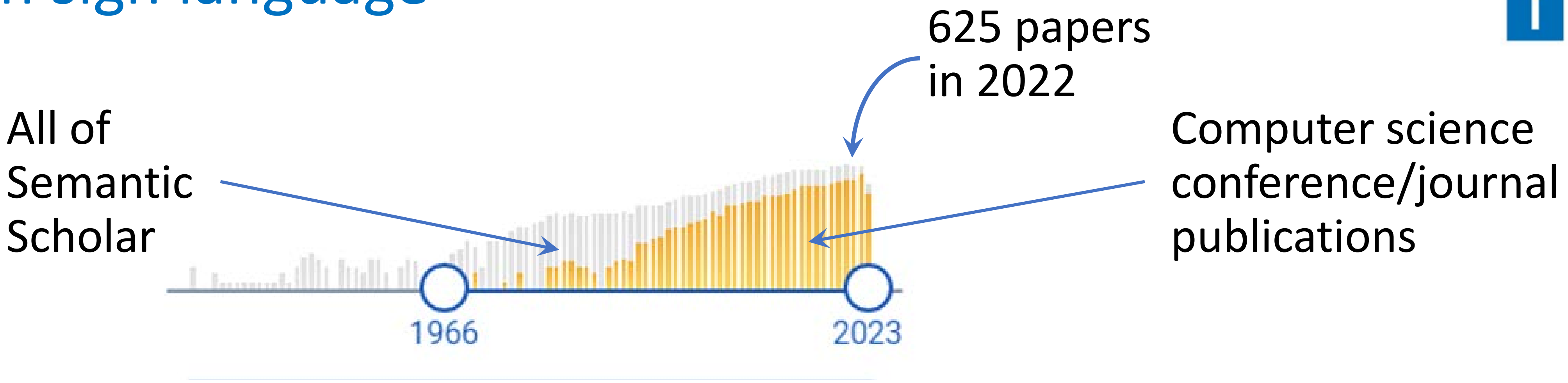
This talk

I. Sign language background

II. Sign language understanding research: Data and tasks

III. Toward sign language understanding “in the wild”: Case studies from a TTIC/U. Chicago collaboration

Research on sign language



- Dominated by computer vision and image processing conference papers
- ACL Anthology: 146 papers
- ISCA Speech archive: 44 papers
- IEEE signal processing conferences + journals: 219 papers
- ICASSP 2023: 3 papers

Sign language understanding tasks

Isolated sign classification

Input: Video clip of a single sign

Output: “read”
(gloss label)

Dataset	Vocab. size	# Signers	# Videos
Purdue RVL-SLLL (Wilbur & Kak 2006)	39	14	546
RWTH-BOSTON-50 (Zahedi et al. 2005)	50	3	483
Boston ASLLVD (Athitsos et al. 2008)	2.742	6	9.794
MS-ASL (Joze & Koller 2018)	1,000	222	25,513
WLASL2000 (Li et al. 2020)	2,000	119	21,083

Sign language understanding tasks

Isolated sign classification

Input: Video clip of a single sign

Output: “read”
(gloss label)

Sign spotting / keyword search

Input: Video clip of a signed utterance + query keyword

Query: “steal”

Output: yes / no

Sign language understanding tasks

Fingerspelling recognition

Input: Fingerspelling Video clip

Output: P-I-R-A-T-E-S

Fingerspelling detection

Input: Raw ASL Video

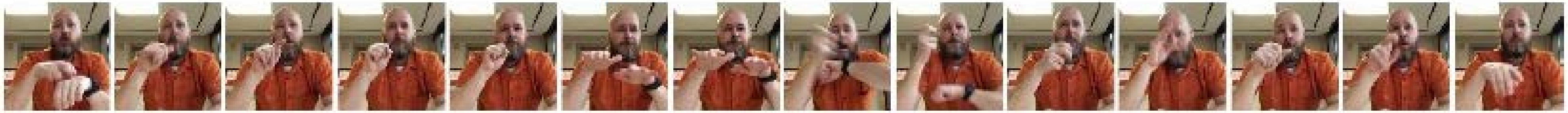
Fingerspelling

Fingerspelling

Output

Sign language understanding tasks

Input: Raw ASL Video



Continuous sign language recognition

Output: P-I-R-A-T-E-S MOVE-FURTIVELY STEAL Point BOY P-A-T-R-I-C-K

Sign language translation

Output: Moving furtively, pirates steal the boy Patrick.

Sign language translation datasets

	Source	Language	Vocab. size	# hours	# signers
Purdue RVL-SLLL (Wilbur et al. 2006)	Lab	ASL	104	-	14
Boston 104 (Dreuw et al. 2007)	Lab	ASL	103	<1	3
Phoenix-2014T (Camgoz et al. 2018)	TV	DGS	3,000	11	9
KETI (Ko et al. 2019)	Lab	KSL	419	28	14
CSL Daily (Zhou et al. 2021)	Lab	CSL	2,000	23	10
SWISSTXT-News (Camgoz et al. 2021)	TV	DSGS	10,000	10	-
BOBSL (Albanie et al. 2021)	TV	BSL	78,000	1467	39
How2Sign (Duarte et al. 2021)	Lab	ASL	16,000	80	11
OpenASL (Shi et al. 2022)	Web	ASL	33,000	288	~220

- Most datasets include both glosses and translations
- Almost all are *interpreted* sign language in a studio setting
- Until very recently, all datasets < 100 hours and < 20 signers

Sign language understanding tasks: How are we doing?

WLASL isolated sign recognition

- Best results obtained with pose tracking model

Method		Top-1 accuracy (%)	Top-5 accuracy (%)
TRN	(Zhou et al. 2018)	49.3	77.9
SL-GCN	(Jiang et al. 2021)	71.0	91.4
Dafnis et al.	(Dafnis et al. 2022)	77.4	94.5

Sign language understanding tasks: How are we doing?

Phoenix-2014T DGS → German translation, using gloss annotations

Method	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
SL-Transf. (Camgoz et al. 2020)	-	46.6	33.7	26.2	21.3
VL-Transfer (Chen et al. 2022)	**52.7**	**54.0**	**41.8**	33.8	28.4
SLTUnet (Zhang et al. 2023)	52.1	52.9	**41.8**	**34.0**	**28.5**

But without gloss annotations…

Method	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GASLT (Yin et al. 2023)	39.9	39.1	26.7	21.9	15.7

This talk

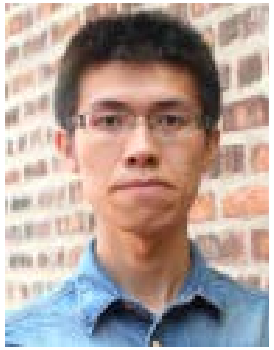
I. Sign language background

II. Sign language understanding from video: Data and tasks

III. Toward sign language understanding "in the wild": Case studies from a TTIC/U. Chicago collaboration





B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu, "Open-domain sign language translation learned from online video," EMNLP 2022

B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu, "Searching for fingerspelled content in American Sign Language," ACL 2022

B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling detection in American Sign Language," CVPR 2021

B. Shi, A. Martinez Del Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling recognition in the wild with iterative visual attention," ICCV 2019

B. Shi, A. Martinez Del Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu, "American Sign Language fingerspelling recognition in the wild," SLT 2018

The evolution of sign language research data

American Sign Language (ASL)
Individual signs, short discourses

Purdue RVL-SLLL
(Wilbur et al., 2006)

ASL
Individual signs

ASLLVD
(Athitsos et al., 2008)

ASL
Fingerspelling sequences

ChicagoFSVid
(Kim et al., 2017)

German Sign Language
Interpreted weather broadcasts

RWTH-Phoenix
(Camgoz et al., 2018)

Korean Sign Language
Individual signs, sentences

KETI
(Ko et al., 2018)

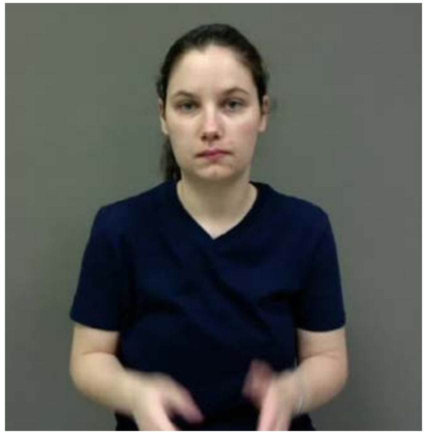
year

The evolution of sign language data

ChicagoFSWild(+)
(Shi et al., 2018-2019)

OpenASL
(Shi et al., 2022)

ASL



MS-ASL
(Joze & Koller, 2019)

ASL

How2Sign
(Duarte et al., 2021)

British Sign Language (BSL)

Content4All
(Camgöz et al., 2021)

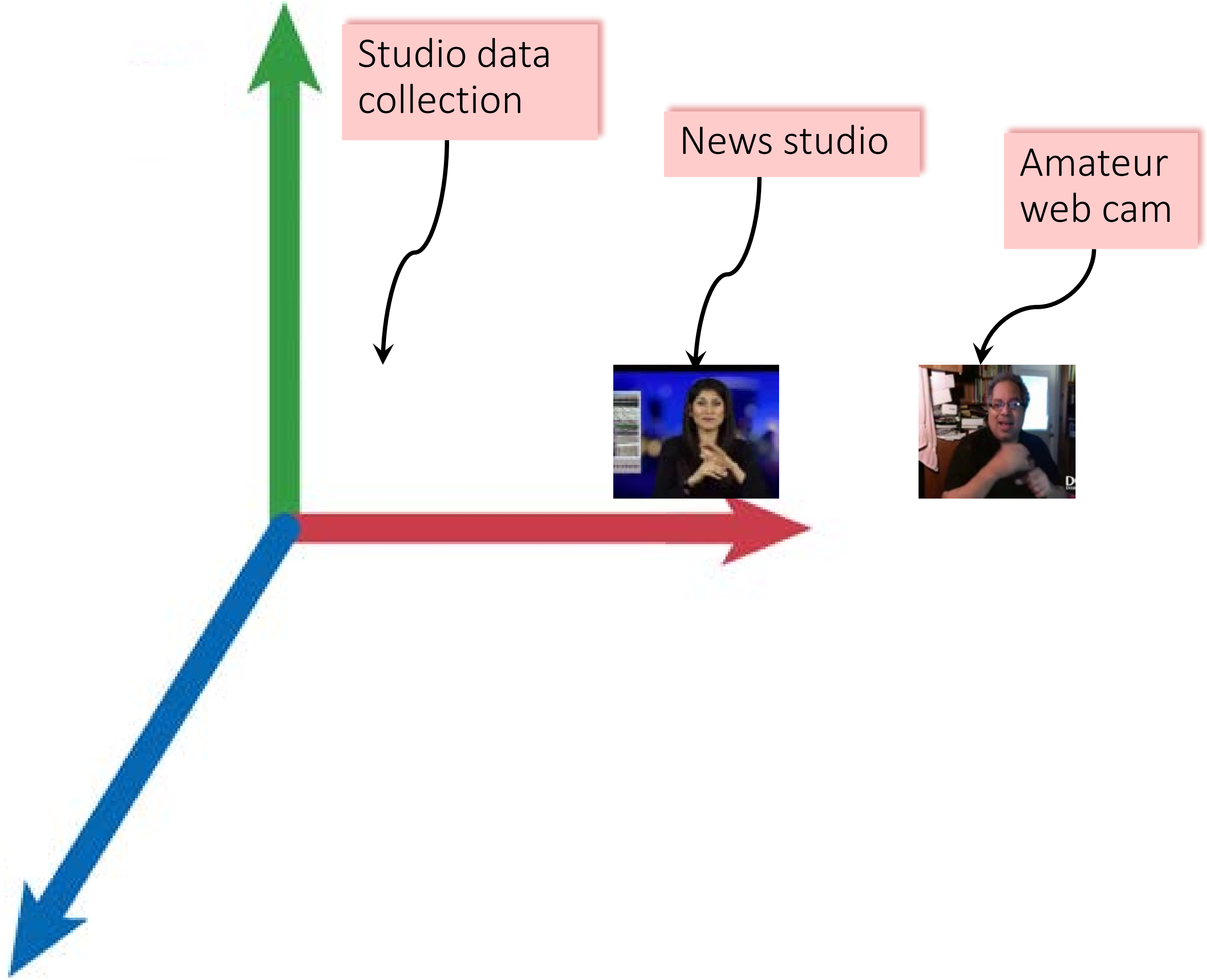
BSL

BOBSL
(Albanie et al., 2021)

year

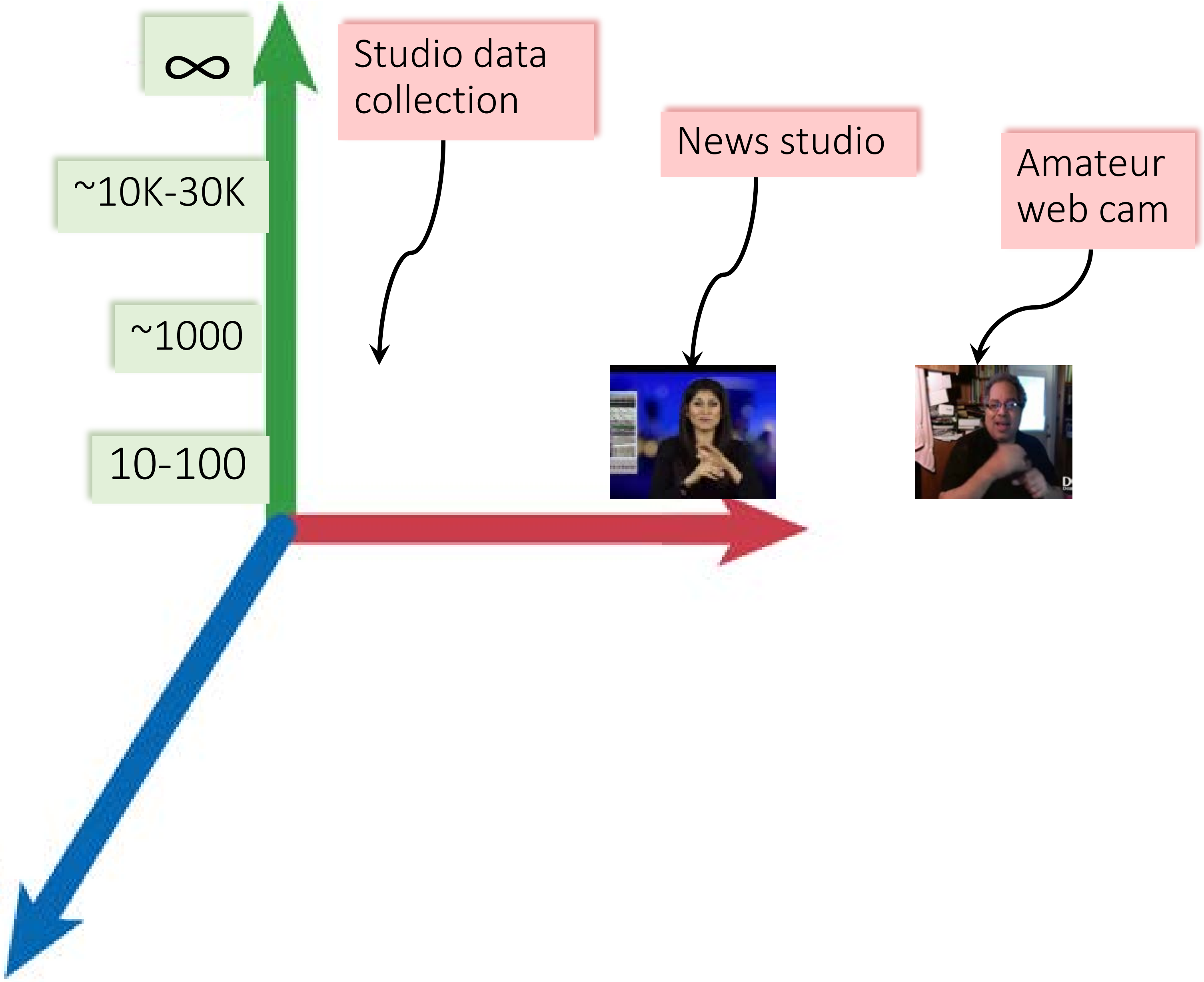
Sign language in the real world: Dimensions of variation

- Visual variability



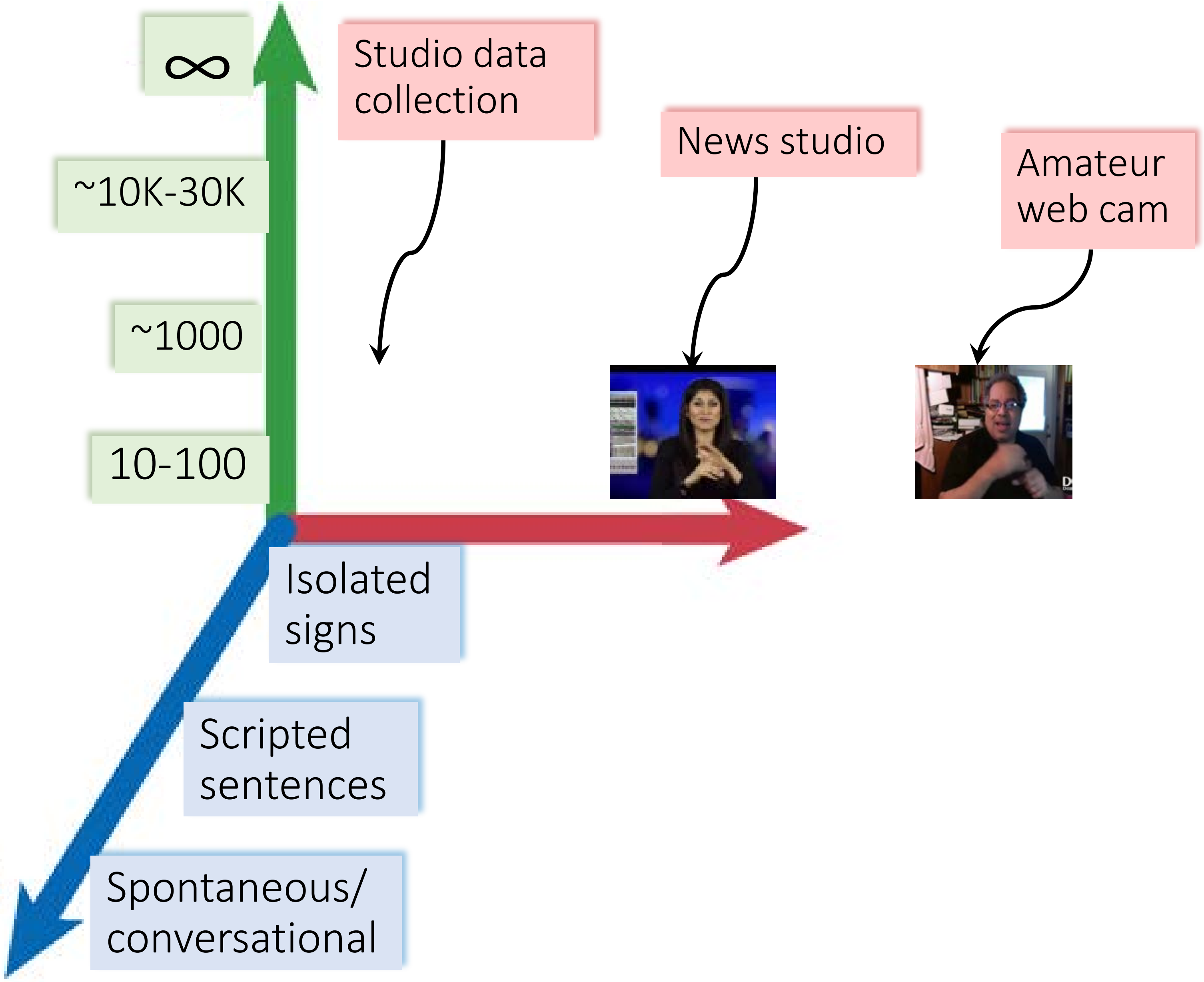
Sign language in the real world: Dimensions of variation

- Visual variability
- Vocabulary size



Sign language in the real world: Dimensions of variation

- Visual variability
- Vocabulary size
- Linguistic complexity
- Other dimensions:
  - Number of signers
  - Interpreted vs. not
  - …



Our goals

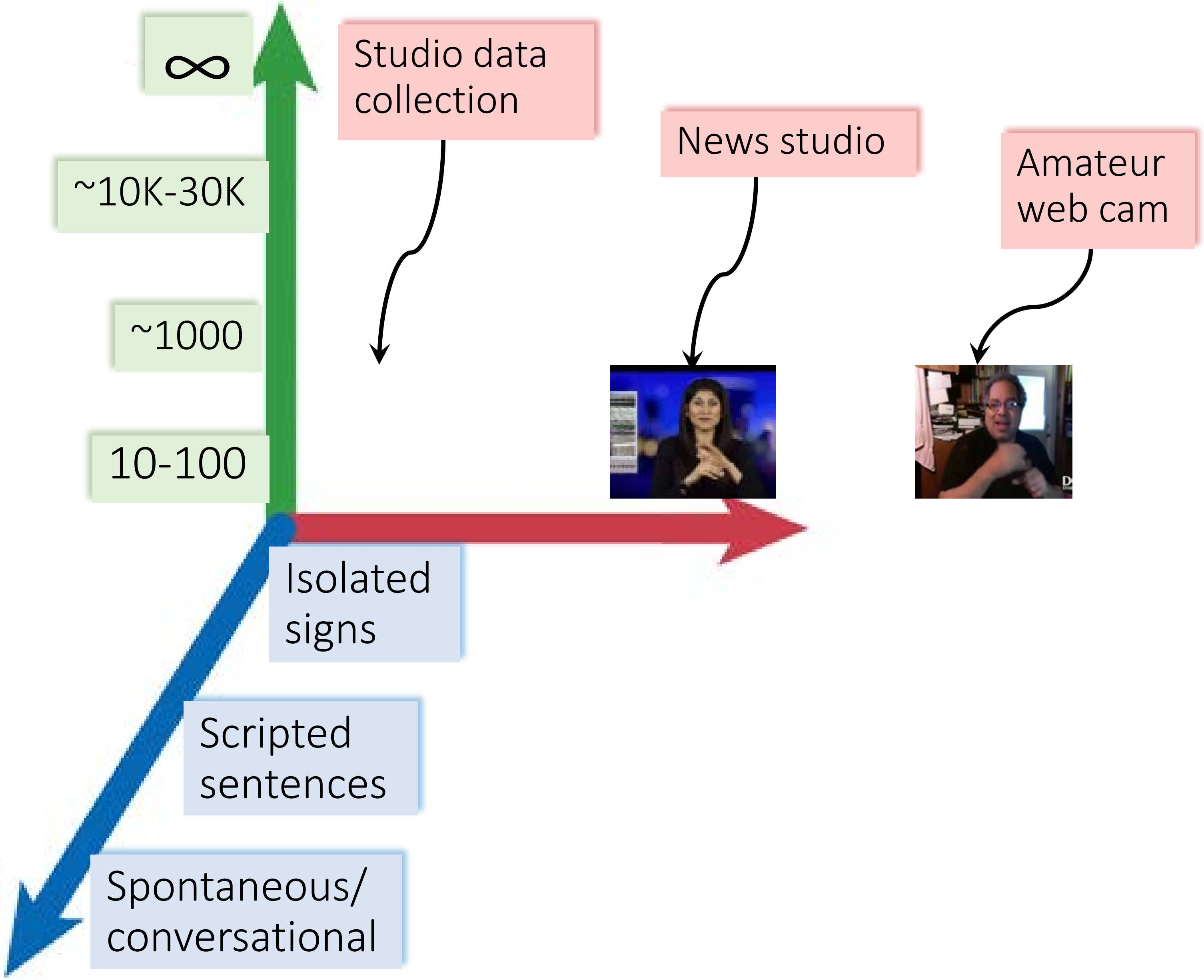
Practical goal: Sign language understanding that is

- Open-domain/vocabulary
- Robust to visual variability
- Signer-independent

… for American Sign Language (for now)

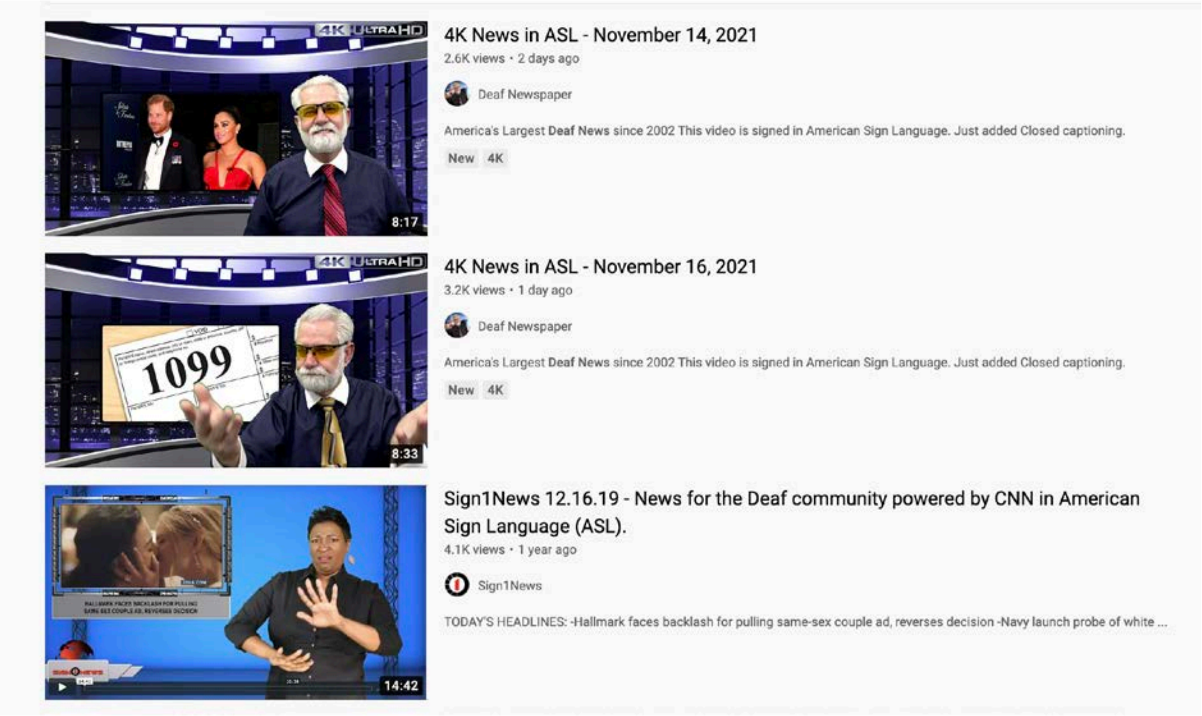
Technical challenges

- Failure of pose trackers, hand detectors, etc.
- Frequent fingerspelling
- No gloss annotations



Data collection from online ASL media

- Large number of signers, open-domain, natural (not interpreted) sign language
- Large visual variability: Lighting, angle, motion blur
- Often, high-quality English captions aligned roughly at the sentence level
- Lots of fingerspelling
- Note: Like other recent web data collections, we don't distribute the videos, only URLs + annotations



ChicagoFSWild

The first real-world ASL fingerspelling dataset

- Sites: YouTube, DeafVideo.tv, aslized.org
- Formats: Vlogs, talks, interviews, …
- Annotated in-house using ELAN
  - Fingerspelling start/end times
  - Fingerspelling transcription

ChicagoFSWild+

- Larger than ChicagoFSWild, and with crowdsourced annotations

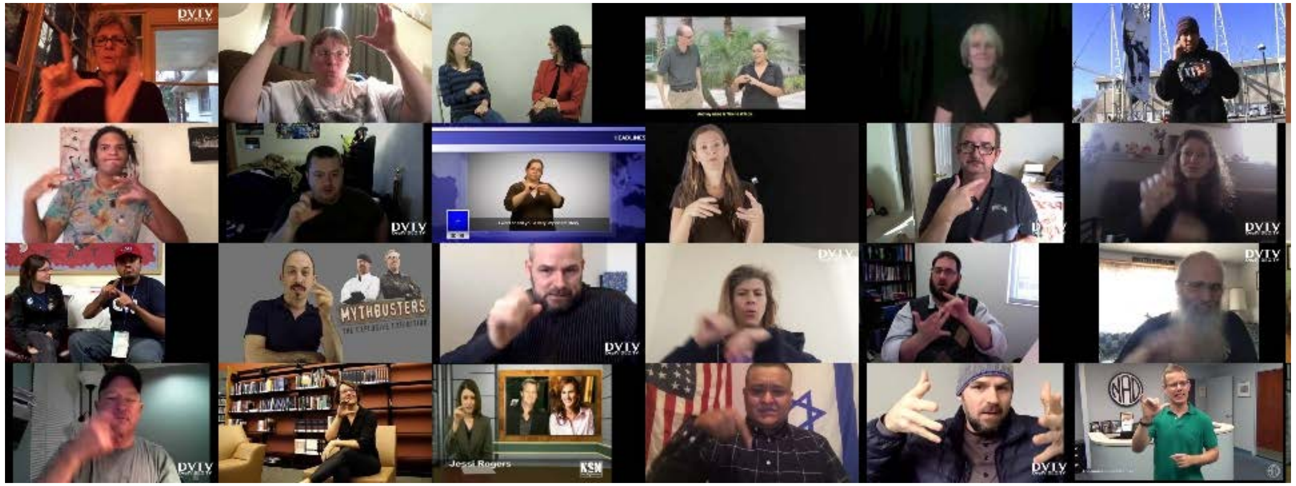
ASL vlogs

Talks

Interviews

Fingerspelling dataset comparison

ChicagoFSWild(+)



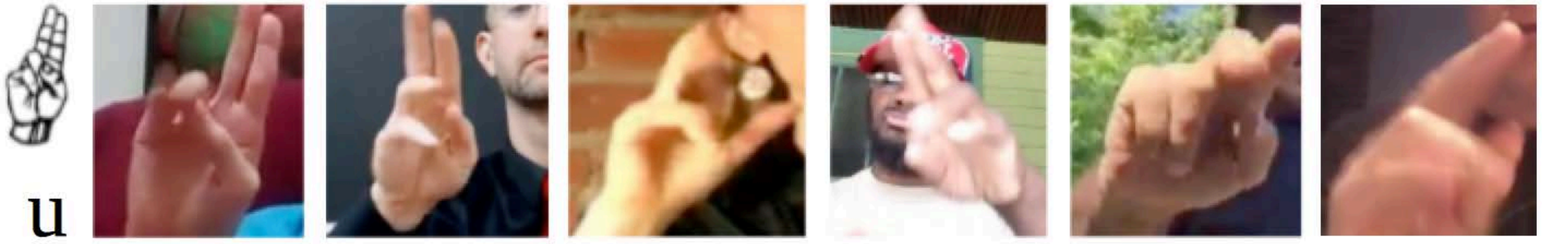
ChicagoFSVid

	Source	Annotation	# sequences	# signers
ChicagoFSVid (Kim et al. 2017)	Studio	In-house	2,400	4
ChicagoFSWild (Shi et al. 2018)	Internet	In-house	7,304	160
ChicagoFSWild+ (Shi et al. 2019)	Internet	Crowdsourced	**55,232**	**260**

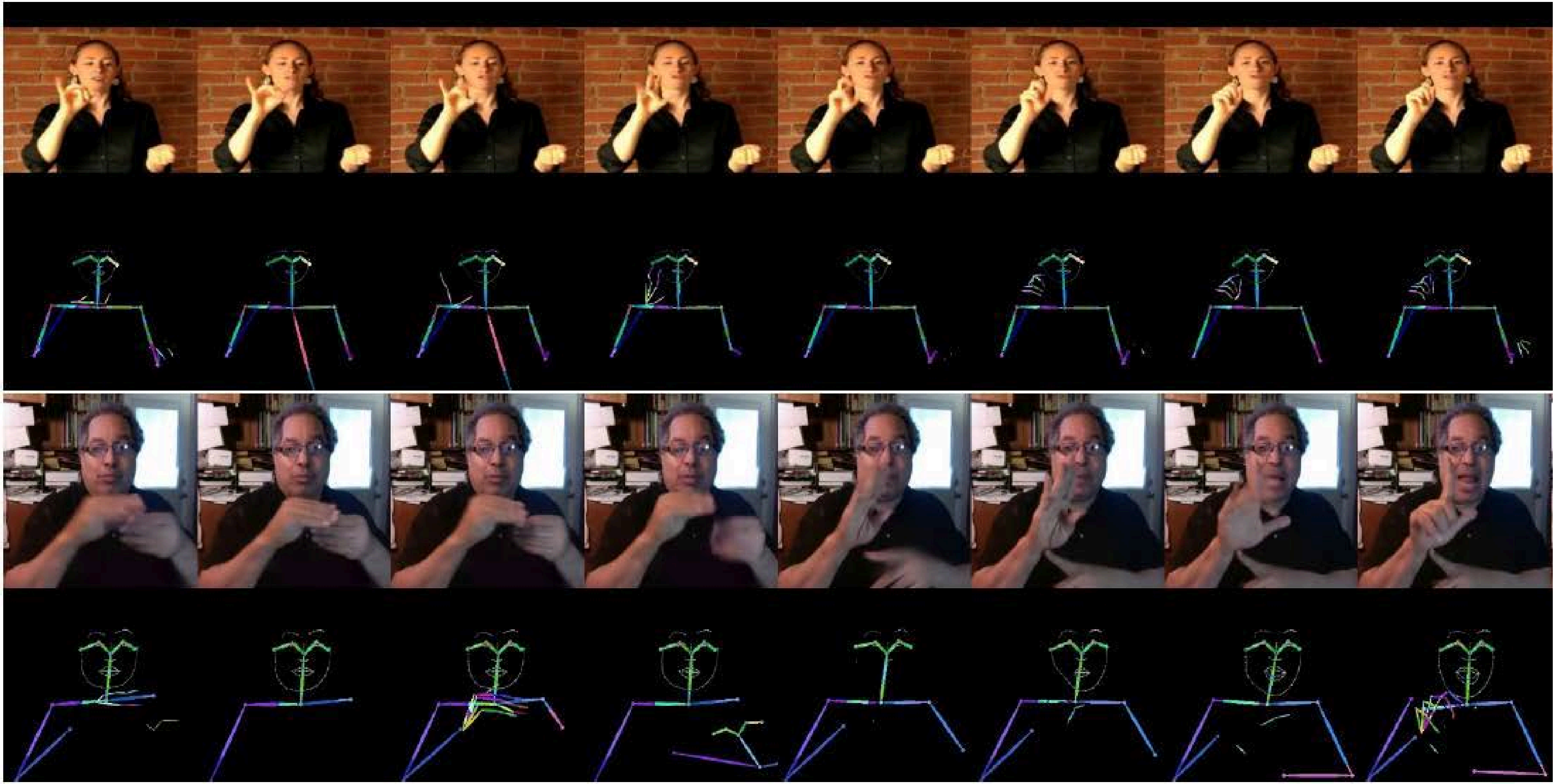
Visual challenges in ChicagoFSWild: Coarticulation

Same signer,
different letters

Same letter,
different signers



Visual challenges in ChicagoFSWild: Pose estimation failure



Task 1: Fingerspelling recognition

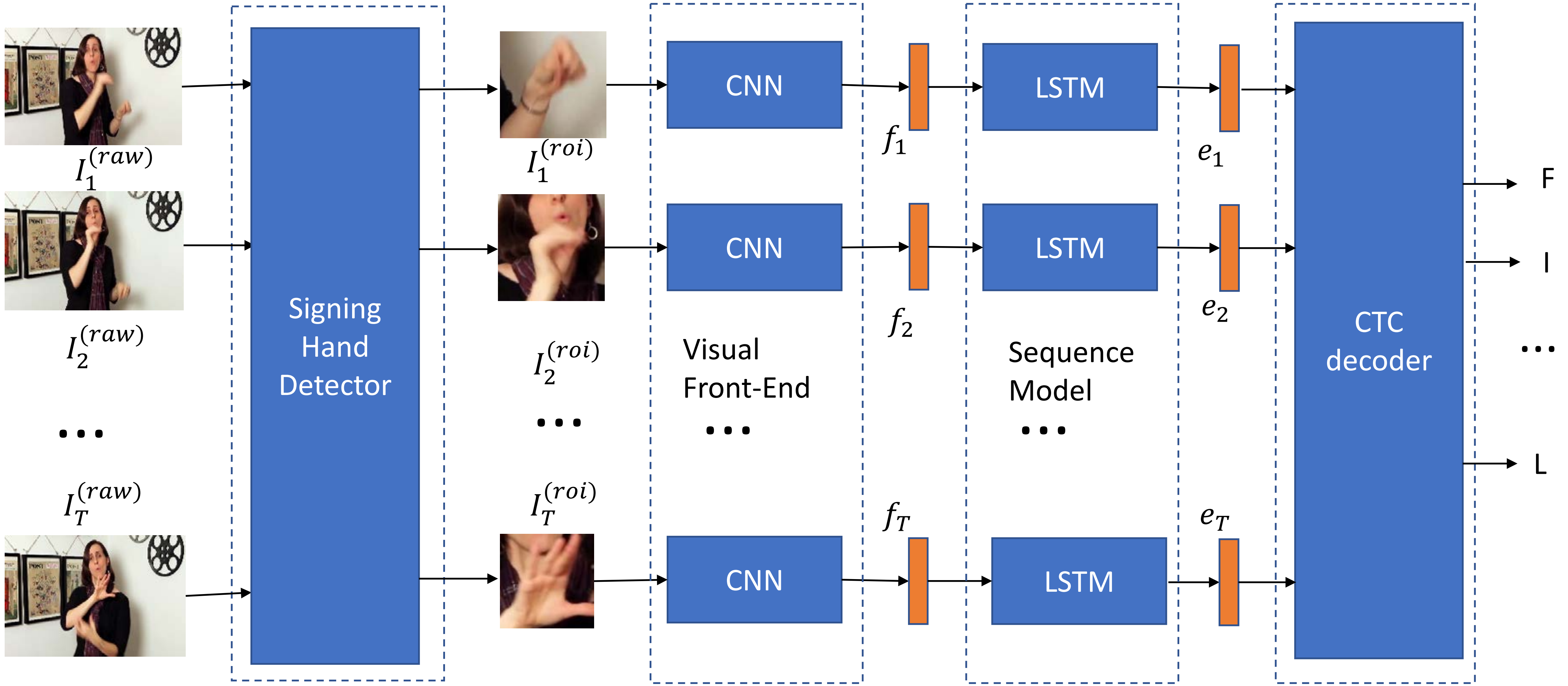
- **Input:** Video clip $I_1, \ldots I_T$ corresponding to a fingerspelling sequence
- **Output:** The letter sequence

(video)

P-I-R-A-T-E-S

Fingerspelling recognition model 1 [Shi et al. 2018]

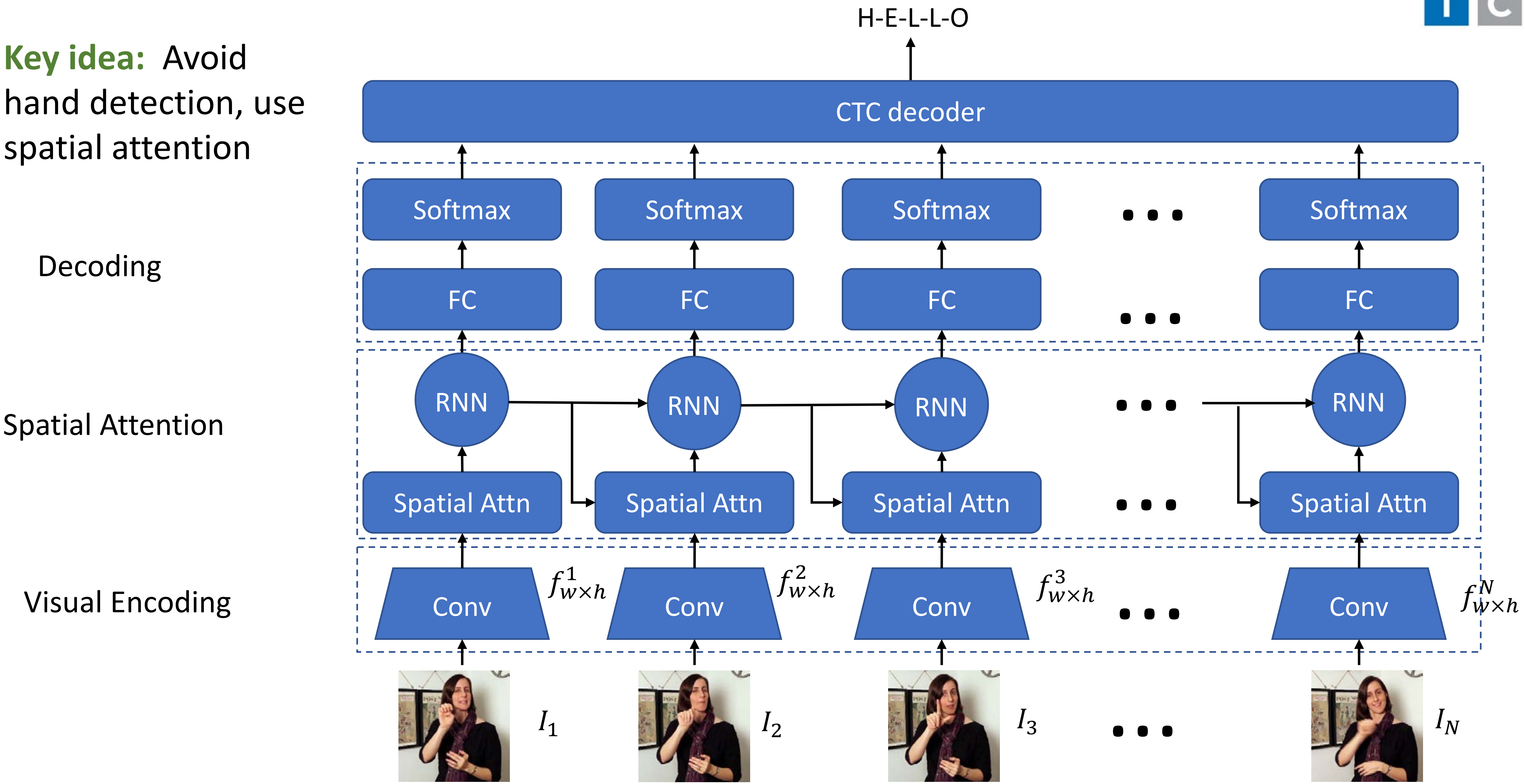
- **Key idea:** Custom-built signing hand detector



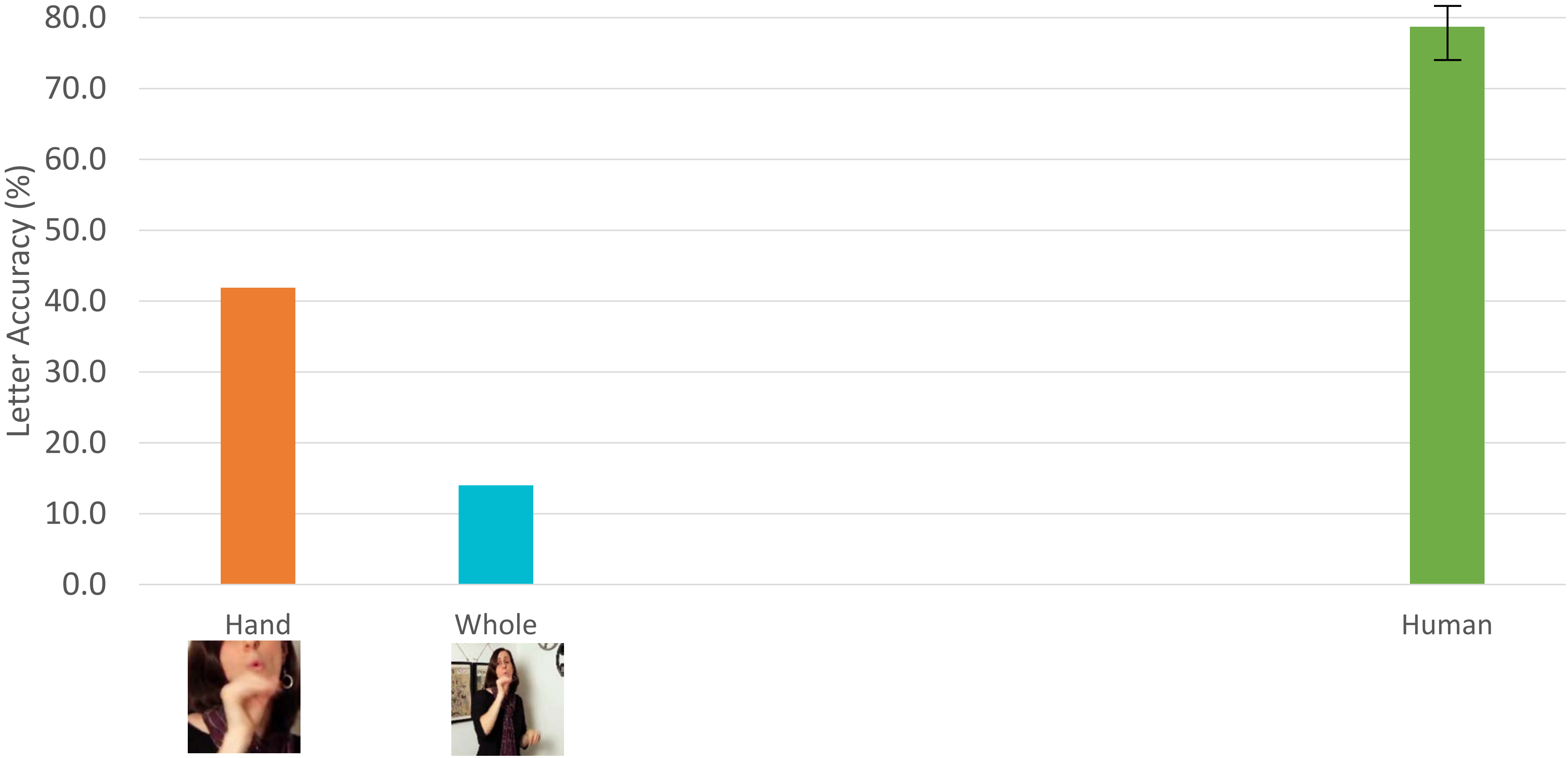
Fingerspelling recognition results

Fingerspelling recognition model 2: End-to-end [Shi et al. 2019]

Key idea: Avoid hand detection, use spatial attention

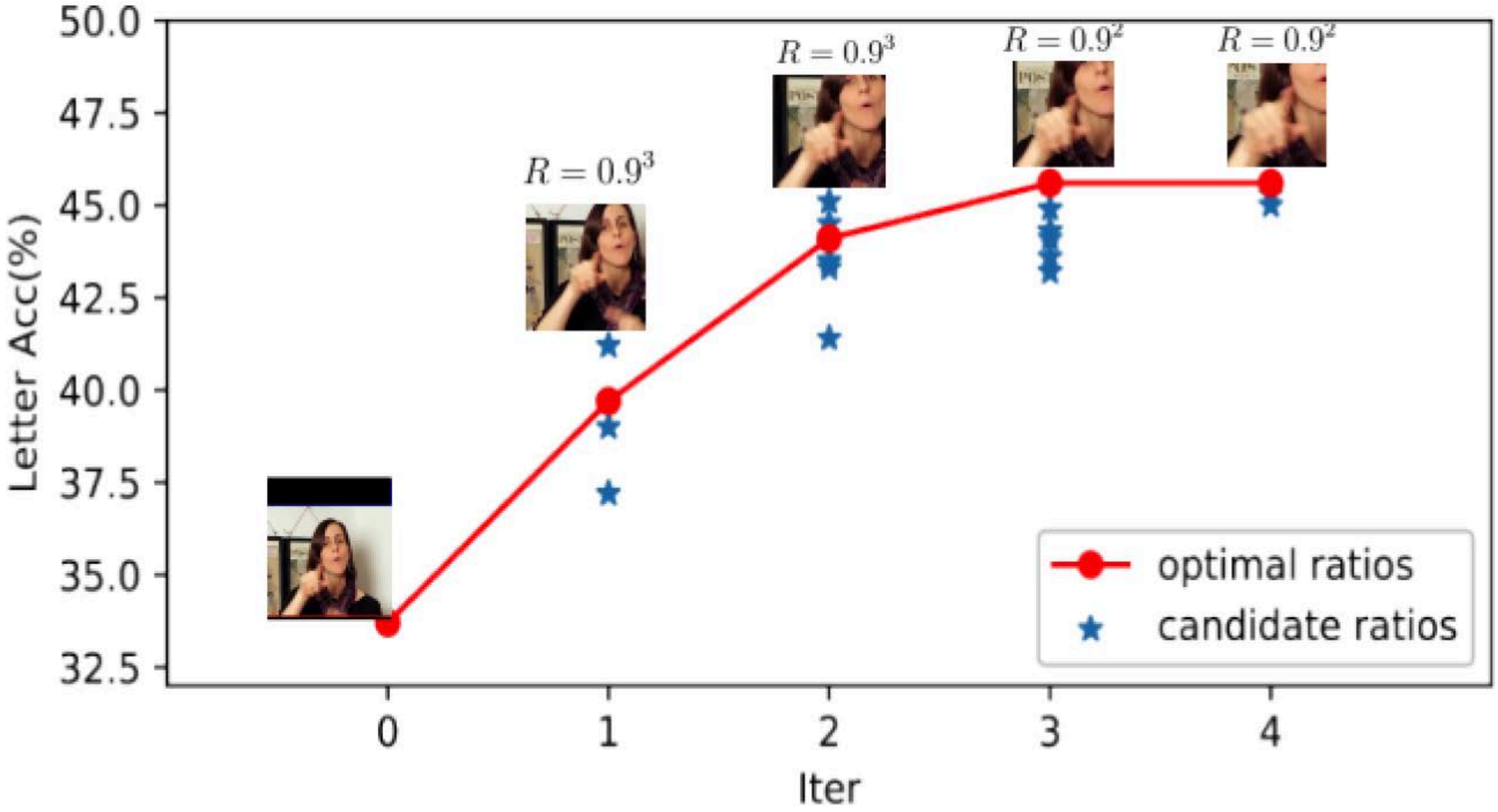


Fingerspelling recognition results



Fingerspelling recognition model 2: Iterative attention [Shi et al. 2019]

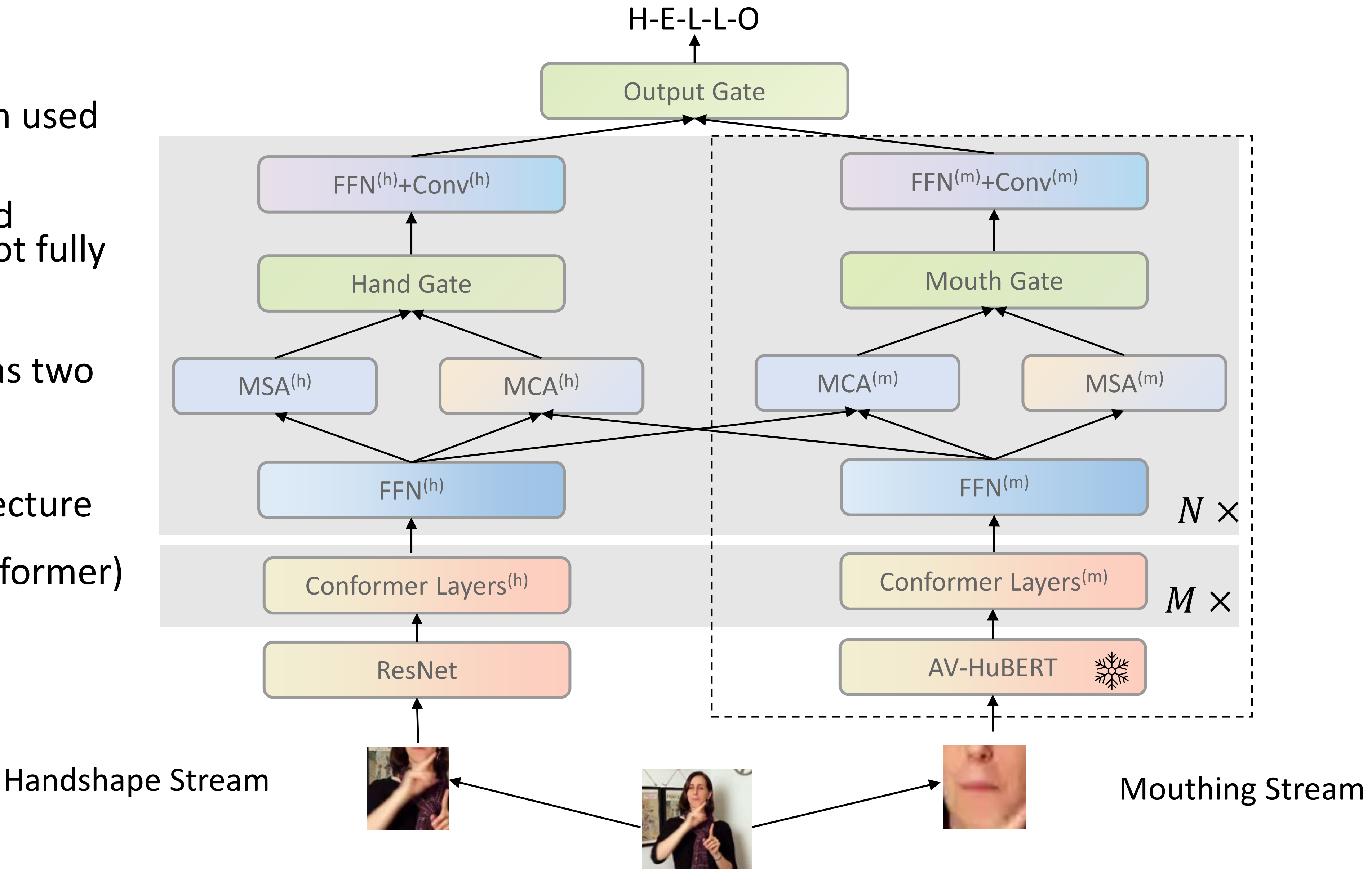
- **Observation**: Fine-grained handshape differences crucial to recognition
- But end-to-end model with full image resolution is computationally expensive
- **Approach**:
  - Zoom in on original image based on attention maps
  - Repeat for multiple iterations, eventually zooming in on signing hand



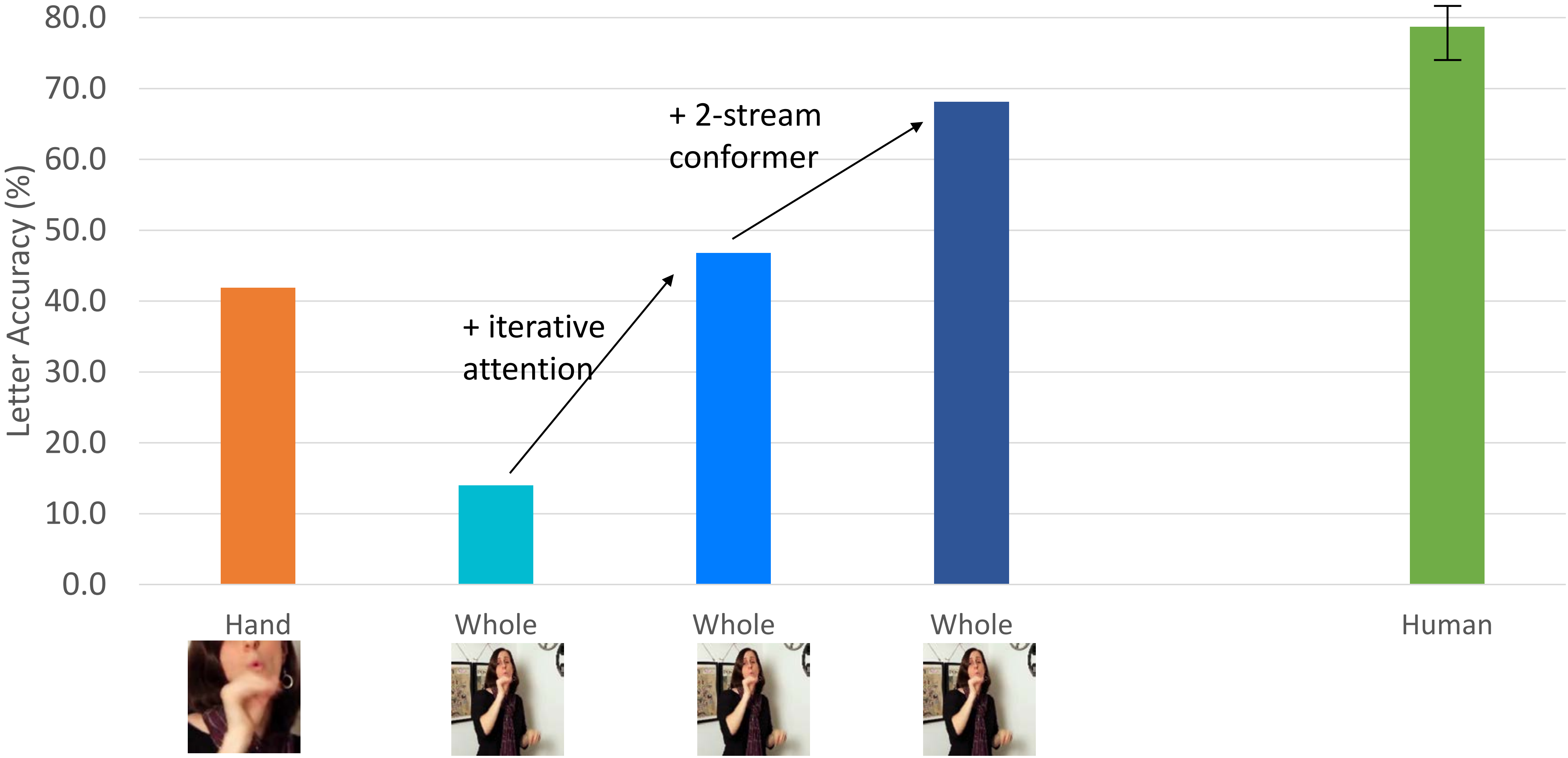
Fingerspelling recognition results

Fingerspelling recognition model 3: Conformer combining hand + mouthing [Shi 2023]

Key ideas:
- Mouthing is often used in fingerspelling
- But mouthing and handshape are not fully synchronized
- So, model them as two separate streams
- (Also, borrow a successful architecture from speech recognition: Conformer)

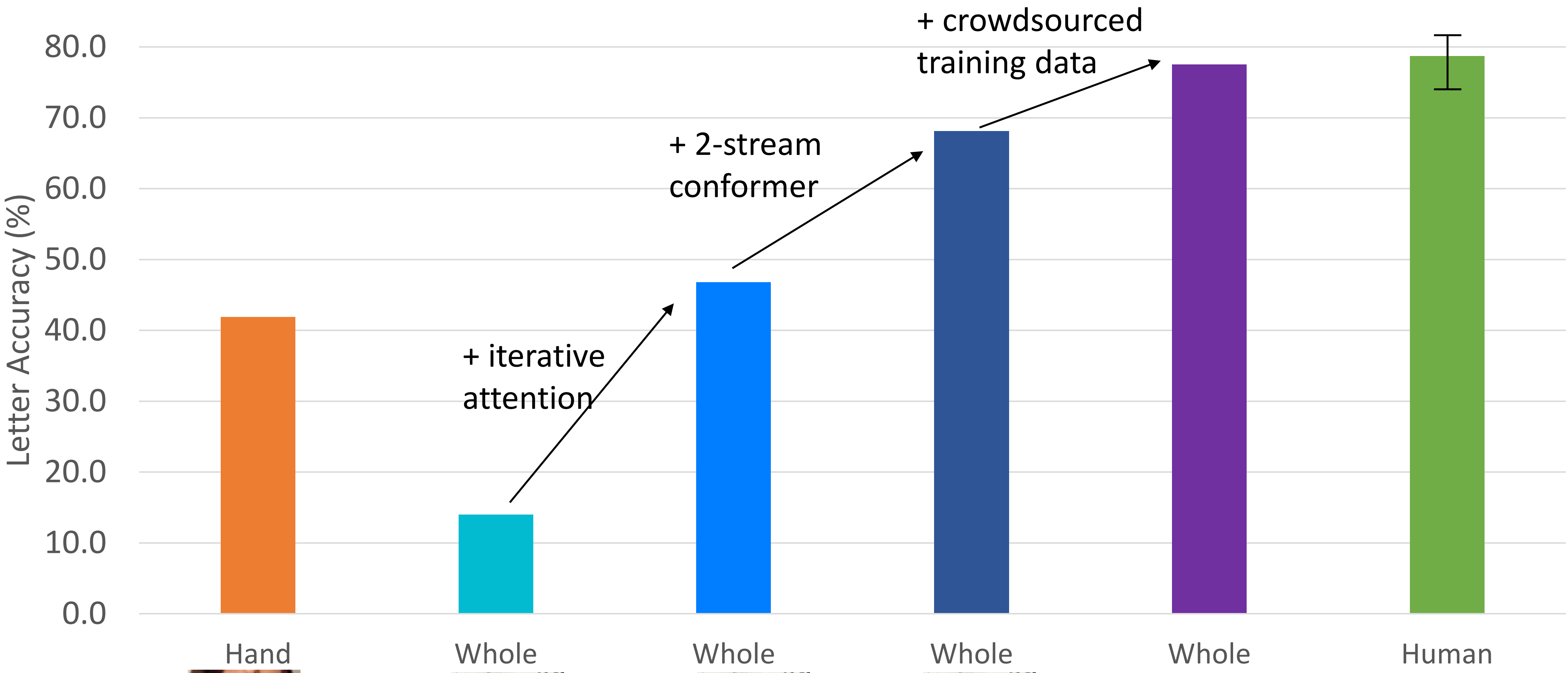


Fingerspelling recognition results



Fingerspelling recognition results

Does noisy crowdsourced training data help?

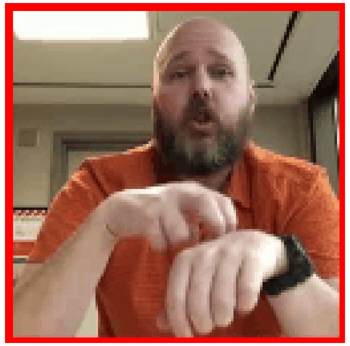


Fingerspelling recognition examples

(video)

(video)

Task 2: Fingerspelling detection [Shi et al. 2021]



(video)

"Moving furtively, pirates steal the boy Patrick."

P-I-R-A-T-E-S

P-A-T-R-I-C-K

Fingerspelling detection example

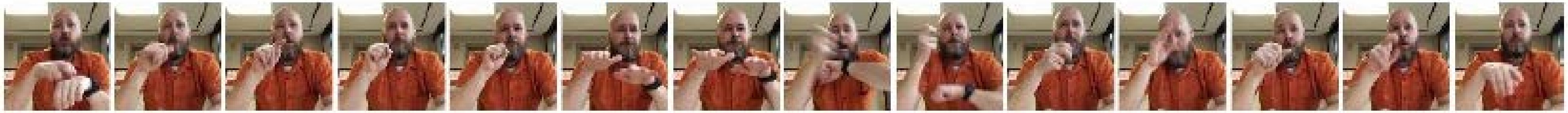
Key ideas

- Detection model trained with multi-task loss combining detection, recognition, and pose
- Pose estimation is a poor feature extractor, but helps as weak supervision
- Outperforms a baseline based on state-of-the-art action recognition

(video)

Task 3: ASL → English translation [Shi et al. 2022]

Input: Raw ASL Video



Output: Moving furtively, pirates steal the boy Patrick.

OpenASL: A real-world ASL translation dataset

Collected from online ASL videos with English captions

- All TheDailyMoth and Sign1News videos through June 2021
- National Association for the Deaf (NAD) YouTube videos: announcements, tips, conversations
- Divided into utterances corresponding to caption sentences

Dev and test sets manually refined by professional captioning service

- Utterance start and end times verified/corrected
- English translations verified/corrected
- Glosses added

OpenASL: Comparison with other translation datasets

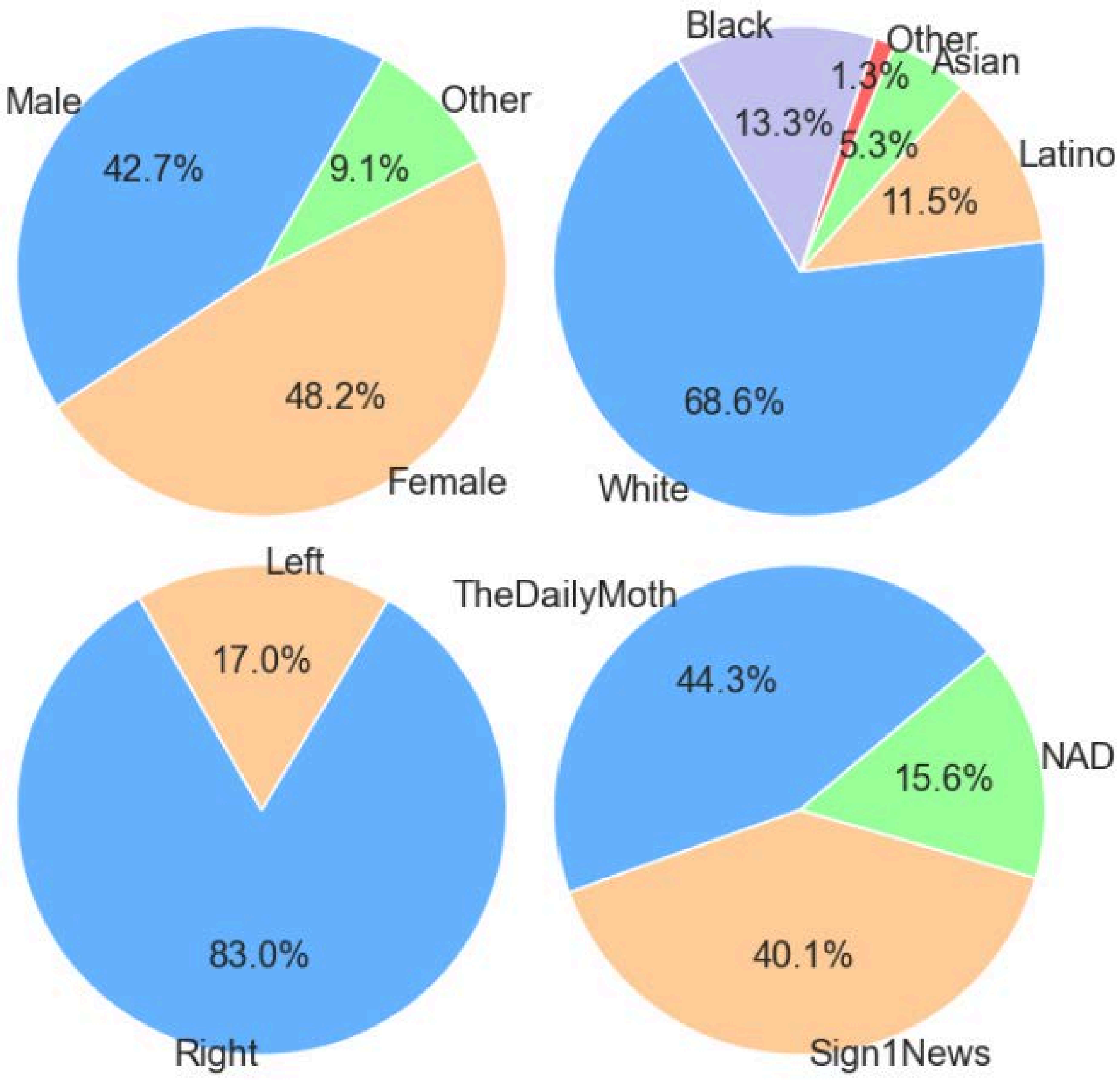
	Source	Language	Vocab. size	# hours	# signers
Purdue RVL-SLLL (Wilbur et al. 2006)	Lab	ASL	104	-	14
Boston 104 (Dreuw et al. 2007)	Lab	ASL	103	<1	3
Phoenix-2014T (Camgoz et al. 2018)	TV	DGS	3,000	11	9
KETI (Ko et al. 2019)	Lab	KSL	419	28	14
CSL Daily (Zhou et al. 2021)	Lab	CSL	2,000	23	10
SWISSTXT-News (Camgoz et al. 2021)	TV	DSGS	10,000	10	-
BOBSL (Albanie et al. 2021)	TV	BSL	78,000	1467	39
How2Sign (Duarte et al. 2021)	Lab	ASL	16,000	80	11
OpenASL (Shi et al. 2022)	Web	ASL	33,000	288	~220

Among ASL datasets, largest vocabulary size, # hours, # signers

Only translation dataset collected from natural (non-interpreted) online video

One downside: No glosses for training set

OpenASL statistics



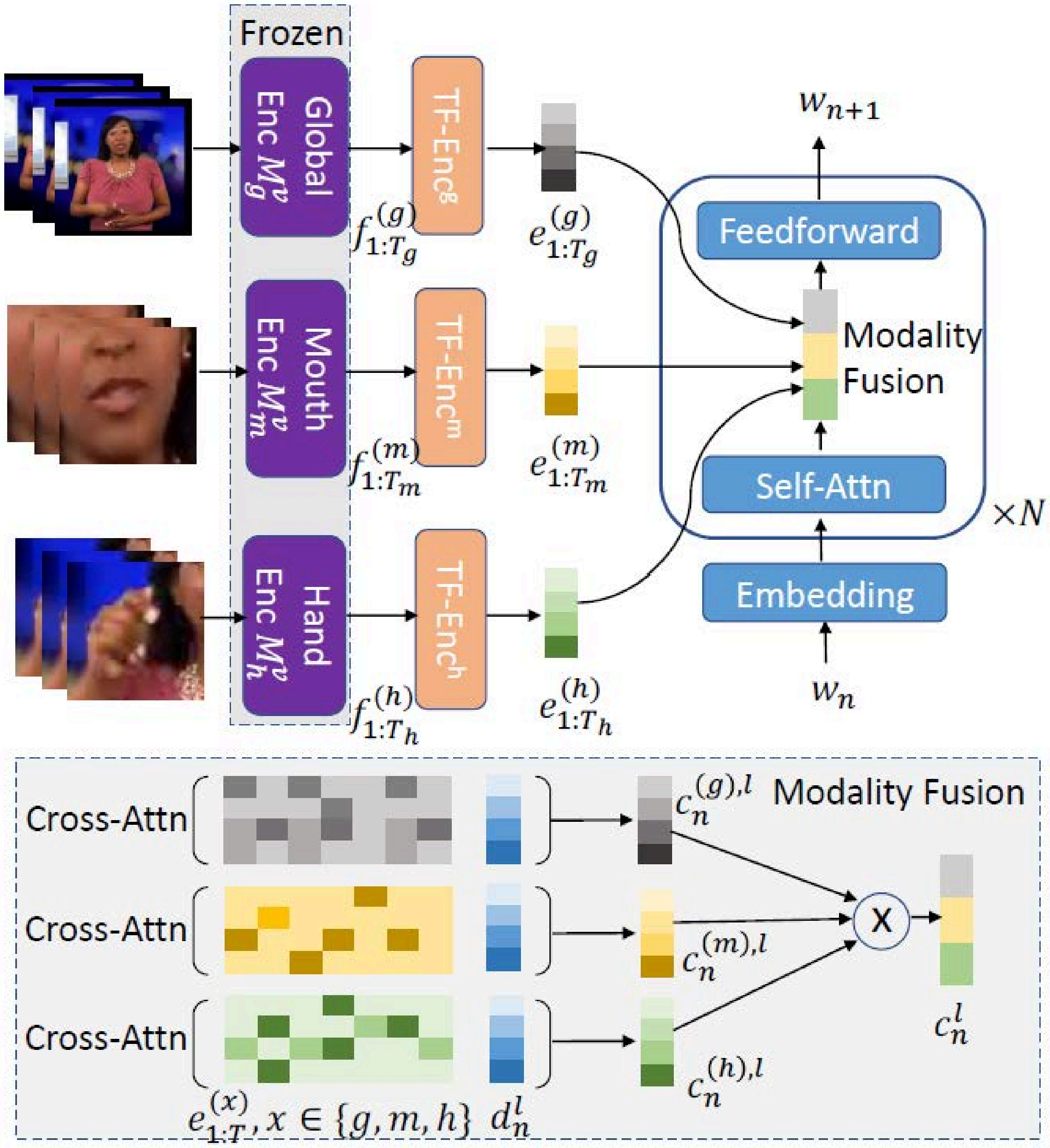
Note:

- Signer characteristics are not ground-truth, but approximate labels from in-house annotators
- >50% of utterances contain fingerspelling

Multi-stream translation model

Key ideas:

- Global + mouth + hand representations with cross-attention
- Low-resource language ➔ rely on pre-training
- Global encoder pre-training
  - Isolated sign classification on WL-ASL (Li et al. 2020)
  - Sign searcn on OpenASL
- Hand encoder pre-trained as a fingerspelling recognizer
- Mouth encoder uses pre-trained AV-HuBERT (Shi et al. 2022)



OpenASL translation performance

Effect of pre-training global model (dev results)

Model	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
iso only	20.91	18.62	11.17	8.24	6.71
+lex	22.44	20.37	12.45	9.15	7.37
+lex+fs	**23.17**	**21.43**	**13.12**	**9.61**	**7.69**

Effect of mouthing and hand features (dev results)

Model	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
global	23.17	21.43	13.12	9.61	7.69
+ m	24.40	22.38	13.75	10.02	7.97
+ m + h	**25.31**	**24.35**	**14.94**	**10.72**	**8.39**

OpenASL translation performance: Final results (so far!)

	DEV						TEST					
Models	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEURT	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEURT
Conv-GRU (Camgoz et al., 2018)†	16.82	16.21	9.15	5.04	3.83	29.00	17.78	15.65	7.55	4.83	3.52	28.80
I3D-transformer	20.91	18.62	11.17	8.24	6.71	32.31	19.83	17.84	9.81	6.76	5.19	30.78
Ours	**25.31**	**24.35**	**14.94**	**10.72**	**8.39**	**34.25**	**24.83**	**23.87**	**14.08**	**9.90**	**7.54**	**34.52**

OpenASL translation performance: Effect of fingerspelling

Utterances with fingerspelling have consistently worse performance, but not by a large margin

Final thoughts

What have we learned about ASL understanding in the real world?

- Online captioned video is a good source of data
- Standard vision components (pose estimation, hand detection) perform poorly, but are useful as additional signals in training/inference
- As in other low-resource tasks, model pre-training is important
- Sign language understanding is not just a combination of existing computer vision + existing NLP

Real-world fingerspelling recognition: Going well!

- Best models match a proficient student of ASL
- Still not matching a native signer

Real-world fingerspelling detection and search: Much work to be done! (Not shown)

Real-world ASL translation: Just getting started!

Many other challenges remain: More pre-training ideas, other sign languages, other tasks, …

Datasets, code:

- https://ttic.edu/livescu/ChicagoFSWild
- https://github.com/chevalierNoir/