Toward Sign Language Video Understanding in the Real World

Karen Livescu
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

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

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

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
Background: Sign language video media

Native (non-interpreted) sign language media

News produced in sign language

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

625 papers in 2022
Sign language understanding tasks

Isolated sign classification

Input: Video clip of a single sign
Output: “read” (gloss label)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Vocab. size</th>
<th># Signers</th>
<th># Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purdue RVL-SLLL (Wilbur &amp; Kak 2006)</td>
<td>39</td>
<td>14</td>
<td>546</td>
</tr>
<tr>
<td>RWTH-BOSTON-50 (Zahedi et al. 2005)</td>
<td>50</td>
<td>3</td>
<td>483</td>
</tr>
<tr>
<td>Boston ASLLVD (Athitsos et al. 2008)</td>
<td>2.742</td>
<td>6</td>
<td>9.794</td>
</tr>
<tr>
<td>MS-ASL (Joze &amp; Koller 2018)</td>
<td>1,000</td>
<td>222</td>
<td>25,513</td>
</tr>
<tr>
<td>WLASL2000 (Li et al. 2020)</td>
<td>2,000</td>
<td>119</td>
<td>21,083</td>
</tr>
</tbody>
</table>
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
Output: Fingerspelling
Fingerspelling
Sign language understanding tasks

Continuous sign language recognition

Sign language translation

Input: Raw ASL Video


Output: Moving furtively, pirates steal the boy Patrick.
Sign language translation datasets

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

- 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 accuracy (%)</th>
<th>Top-5 accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRN</td>
<td>49.3</td>
<td>77.9</td>
</tr>
<tr>
<td>SL-GCN</td>
<td>71.0</td>
<td>91.4</td>
</tr>
<tr>
<td>Dafnis et al.</td>
<td>77.4</td>
<td>94.5</td>
</tr>
</tbody>
</table>
Sign language understanding tasks: How are we doing?

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

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL-Transf.</td>
<td></td>
<td>46.6</td>
<td>33.7</td>
<td>26.2</td>
<td>21.3</td>
</tr>
<tr>
<td>(Camgoz et al. 2020)</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VL-Transfer</td>
<td></td>
<td><strong>52.7</strong></td>
<td><strong>54.0</strong></td>
<td><strong>41.8</strong></td>
<td>33.8</td>
</tr>
<tr>
<td>(Chen et al. 2022)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLTUnet</td>
<td>52.1</td>
<td>52.9</td>
<td><strong>41.8</strong></td>
<td><strong>34.0</strong></td>
<td><strong>28.5</strong></td>
</tr>
<tr>
<td>(Zhang et al. 2023)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But without gloss annotations...

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GASLT</td>
<td>39.9</td>
<td>39.1</td>
<td>26.7</td>
<td>21.9</td>
<td>15.7</td>
</tr>
<tr>
<td>(Yin et al. 2023)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
The evolution of sign language research data

American Sign Language (ASL)
Individual signs, short discourses

ASL
Individual signs

ASL
Fingerspelling sequences

German Sign Language
Interpreted weather broadcasts

Korean Sign Language
Individual signs, sentences

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

ASLLVD
(Athitsos et al., 2008)

ChicagoFSVid
(Kim et al., 2017)

RWTH-Phoenix
(Camgoz et al., 2018)

KETI
(Ko et al., 2018)
The evolution of sign language data

- How2Sign (Duarte et al., 2021)
- MS-ASL (Joze & Koller, 2019)
- ChicagoFSWild(+) (Shi et al., 2018-2019)
- OpenASL (Shi et al., 2022)
- Content4All (Camgöz et al., 2021)
- BOBSL (Albanie et al., 2021)
Sign language in the real world: Dimensions of variation

- Visual variability
Sign language in the real world: Dimensions of variation

- Visual variability
- Vocabulary size

- Studio data collection
- News studio
- Amateur web cam

Dimensions:
- Visual variability:
  - Infinity
- Vocabulary size:
  - ~10K-30K
  - ~1000
  - 10-100
Sign language in the real world: Dimensions of variation

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

Studio data collection

- ~10K-30K
- ~1000
- 10-100

Isolated signs

- Scripted sentences
- Spontaneous/conversational

News studio

Amateur web cam
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
Fingerspelling dataset comparison

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Annotation</th>
<th># sequences</th>
<th># signers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChicagoFSVid</td>
<td>Studio</td>
<td>In-house</td>
<td>2,400</td>
<td>4</td>
</tr>
<tr>
<td>(Kim et al. 2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChicagoFSWild</td>
<td>Internet</td>
<td>In-house</td>
<td>7,304</td>
<td>160</td>
</tr>
<tr>
<td>(Shi et al. 2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChicagoFSWild+</td>
<td>Internet</td>
<td>Crowdsourced</td>
<td>55,232</td>
<td>260</td>
</tr>
<tr>
<td>(Shi et al. 2019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
Key idea: Custom-built signing hand detector
Fingerspelling recognition results

Hand: 40.0%

Human: 80.0%
Fingerspelling recognition model 2: End-to-end [Shi et al. 2019]

**Key idea:** Avoid hand detection, use spatial attention

---

**Decoding**

- CTC decoder
- Softmax
- FC
- RNN
- Spatial Attn

---

**Spatial Attention**

- $f_{w\times h}^1$
- $f_{w\times h}^2$
- $f_{w\times h}^3$
- $f_{w\times h}^N$

---

**Visual Encoding**

- Conv
- $I_1$
- $I_2$
- $I_3$
- $I_N$
Fingerspelling recognition results

Letter Accuracy (%)

- Hand: 40.0%
- Whole: 10.0%
- Human: 70.0%
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

Letter Accuracy (%)

- Hand
- Whole
- Whole + iterative attention
- Human
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

- **Hand**: 40.0%
- **Whole + iterative attention**: 50.0%
- **Whole + 2-stream conformer**: 60.0%
- **Human**: 80.0%

The diagram compares the letter accuracy of different methods and shows how adding attention mechanisms (iterative and 2-stream) improves the recognition rate.
Fingerspelling recognition results

Does noisy crowdsourced training data help?
Fingerspelling recognition examples

Hyp:
Ref: SCIENCE FICTION (video)

Hyp:
Ref: OPEN SOURCE (video)
“Moving furtively, **pirates** steal the boy **Patrick.**”
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
Task 3: ASL $\rightarrow$ 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

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

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 search 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)

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>iso only</td>
<td>20.91</td>
<td>18.62</td>
<td>11.17</td>
<td>8.24</td>
<td>6.71</td>
</tr>
<tr>
<td>+lex</td>
<td>22.44</td>
<td>20.37</td>
<td>12.45</td>
<td>9.15</td>
<td>7.37</td>
</tr>
<tr>
<td>+lex+fs</td>
<td><strong>23.17</strong></td>
<td><strong>21.43</strong></td>
<td><strong>13.12</strong></td>
<td><strong>9.61</strong></td>
<td><strong>7.69</strong></td>
</tr>
</tbody>
</table>

### Effect of mouthing and hand features (dev results)

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>global</td>
<td>23.17</td>
<td>21.43</td>
<td>13.12</td>
<td>9.61</td>
<td>7.69</td>
</tr>
<tr>
<td>+ m</td>
<td>24.40</td>
<td>22.38</td>
<td>13.75</td>
<td>10.02</td>
<td>7.97</td>
</tr>
<tr>
<td>+ m + h</td>
<td><strong>25.31</strong></td>
<td><strong>24.35</strong></td>
<td><strong>14.94</strong></td>
<td><strong>10.72</strong></td>
<td><strong>8.39</strong></td>
</tr>
</tbody>
</table>
# OpenASL translation performance: Final results (so far!)

<table>
<thead>
<tr>
<th>Models</th>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ROUGE</td>
<td>BLEU-1</td>
</tr>
<tr>
<td>Conv-GRU (Camgoz et al., 2018)†</td>
<td>16.82</td>
<td>16.21</td>
</tr>
<tr>
<td>I3D-transformer</td>
<td>20.91</td>
<td>18.62</td>
</tr>
<tr>
<td>Ours</td>
<td>25.31</td>
<td>24.35</td>
</tr>
</tbody>
</table>
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://ttic.edu/livescu/ChicagoFSWild)
- [https://github.com/chevalierNoir/](https://github.com/chevalierNoir/