**Motivation**

Signs are defined by 5 parameters: handshape, orientation, location, movement, facial expression.

Human’s ability to distinguish between sign minimal pairs with only handshape differences is limited. [1]

A handshape minimal pair in ASL. Difficult to distinguish when signers sign rapidly.

**Sign Language Recognition (SLR) system**

- video encoder: (1) visual encoder – spatial  
  (2) sequence encoder – temporal
- predictor: alignment module

**Contribution:**
1. Extend an existing dataset with handshapes.  
2. Propose two handshape-inclusive SLR systems.

**Challenges:**
Signs are interpreted as a cohesive whole.

**Solutions:**
Incorporation of phonological features into SLR systems.

**References**


**PHOENIX14T-HS: The Handshape-Extended Dataset**

<table>
<thead>
<tr>
<th>Sign</th>
<th>MORGEN</th>
<th>TEMPERATUR</th>
<th>ACHT</th>
<th>BIS</th>
<th>DREIZEHN</th>
<th>MAXIMAL</th>
<th>DREIZEHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng</td>
<td>(TOMORROW)</td>
<td>(TEMPERATURE)</td>
<td>(EIGHT)</td>
<td>(TO)</td>
<td>(THIRTEEN)</td>
<td>(MAXIMUM)</td>
<td>(THIRTEEN)</td>
</tr>
</tbody>
</table>

**Glosses**
- Handshape
- Right Handshape
- Left Handshape

**Dataset info:**
German sign language for Weather forecast [2]; train/dev/test: 7,096/519/642; vocab: 1,066; average gloss length: 7.8

**Annotation Process:**
1. Look up the SignWriting dictionary (a publicly accessible, user-edited sign language dataset, with handshape annotations).
2. Manually label handshapes for signs not present in SignWriting. (1/3 vocab; the author is the annotator)

**Comparison with SOTA (WER)**

<table>
<thead>
<tr>
<th>Method</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-LSTM (Koller et al., 2019)*</td>
<td>22.1</td>
<td>24.1</td>
</tr>
<tr>
<td>SPL (Niu and Mak, 2020)</td>
<td>25.1</td>
<td>26.1</td>
</tr>
<tr>
<td>FCN (Cheng et al., 2020)</td>
<td>23.3</td>
<td>25.3</td>
</tr>
<tr>
<td>Joint-SLRT (Camgoz et al., 2020)</td>
<td>24.6</td>
<td>24.5</td>
</tr>
<tr>
<td>CMA (Papoutsis et al., 2020)*</td>
<td>23.9</td>
<td>24.0</td>
</tr>
<tr>
<td>SignSLT (Zhou et al., 2021a)</td>
<td>22.7</td>
<td>23.9</td>
</tr>
<tr>
<td>MMLTLB (Chen et al., 2021)</td>
<td>21.9</td>
<td>22.5</td>
</tr>
<tr>
<td>SMKD (Hao et al., 2021)</td>
<td>20.8</td>
<td>22.4</td>
</tr>
<tr>
<td>HS-SLWrecker</td>
<td>20.3</td>
<td>21.8</td>
</tr>
</tbody>
</table>

**Conclusions from Ablation Studies**

1. baseline 23.69 > HS-SLRT 21.8 WER
2. Model II is better than Model I
3. incorporating both hands is not better than the only right hand
4. pretraining is helpful
5. freezing parts of S3D is helpful
6. Adding CE-loss is helpful

**Conclusions**

2. It is beneficial to have handshape annotations.

**Future directions:** (1) multi-modal (2) contrastive learning

**Handshape-inclusive SLR systems**

- Model I.
  - Sign Video
  - S3D
  - Head → Predictor → Glosess
  - Head → Predictor → Left HS
  - Head → Predictor → Right HS
  - Sign Video
  - S3D
  - Head → Predictor → Glosess
  - Head → Predictor → Left HS
  - Head → Predictor → Right HS

- Model II.
  - Sign Video
  - S3D
  - Head → Joint Head → Predictor → Glosess
  - Head → Joint Head → Predictor → Left HS
  - Head → Joint Head → Predictor → Right HS

**Baseline System** [3] (w/o handshape)