Deep Learning of Mouth Shapes for Sign Language

Oscar Koller¹, Hermann Ney¹ and Richard Bowden²

¹Human Language Technology and Pattern Recognition, RWTH Aachen University, Germany
²Center for Vision, Speech and Signal Processing, University of Surrey, UK

Sign Language Basics

- Sign languages are natural languages
- Visual languages
- Grammar differs from spoken language
- Not international
- No standardised writing system
- Not yet legally recognised in many countries
- Convey information by [1]:
  - Manual parameters
  - Non-manual parameters

Sign Languages are natural languages
But mouthings required for full understanding
HMM framework (in bakis structure) with 3 GMM single density states

Annotations:
Sign language interpreted weather forecast
Research focused on hands only

Solution

- No mouthing annotations in most corpora
- But mouthings required for full understanding
- Historically, sign language recognition research focused on hands only

Approach

- Find sequence of visemes \(x^n_t \rightarrow x^n_1\) belonging to sequence of mouthings \(m^n \rightarrow m_1, \ldots, m_n\), where the sequence of features \(x^n_t \rightarrow x^n_1, \ldots, x^n_T\) best matches the viseme models.
- Maximise posterior prob. \(p(x^n_t | m^n)\) with EM over all viseme sequences for given gloss sequence.
- \(p(m^n):\) pronunciation probability (in the pronunciation lexicon) for a chosen mouthing.
- HMM framework (in bakis structure) with 3 GMM single density states
- Shared transition probabilities for all visemes, except the ‘no-mouthing’ model
- Using pretrained GoogleNet CNN (22 layers) to model \(p(v|x)\) and employ \(p(x^n_t | m^n) \propto p(v|x^n_t) \cdot p(v)\).

Corpora: RWTH-PHOENIX-Weather

- Sign language interpreted weather forecast
- Recorded from public TV 2009-2013
- German sign language (OGS)

Annotations:

Conclusions

- Improve AAM features by 8% 47.1% → 55.7%
- No feature preproccesing needed
- Equal learning rate across CNN is beneficial, even when fine-tuning
- Websly supervised learning requires shuffling

Future Work

- Integrate into sign language recognition
- Combine with other modalities
- Model mouth gestures

References


Results

- Two tasks:
  1) alignment (with language model)
  2) classification
- DNNs outperform AAM based approach.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAM</td>
<td>64.5</td>
<td>44.2</td>
</tr>
<tr>
<td>CNN Viseme</td>
<td>55.7</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Table: Accuracies in [%] comparing CNN to AAM.