Sign Language Detection “in the Wild” with Recurrent Neural Networks

Mark Borg  
Kenneth P. Camilleri  
Systems and Control Engineering,  
University of Malta

IEEE 2019 ICASSP  
International Conference on Acoustics, Speech and Signal Processing
Sign Languages

• Visual languages
• Multi-modal
• Concurrent modalities

• Articulators:
  • Manual
    • Hand motion
    • Hand shapes
    • Place of articulation
  • Non-Manual
    • Mouth patterns
    • Facial Expressions
    • Body posture

American Sign Language (ASL): PERSON WHATEVER IX-they JUDGE IX1 REALLY WASTE, MINUTE [shook-head] INSTEAD-OF MINUTE IX WHAT-conj LOVE ACCEPT WHO

English equivalent: For every minute we judge, we have squandered a minute we could have used to accept and love someone.

Source: HandSpeak
Sign Language technologies

- Web Video Repositories
- Sign Language Detection
- Sign Language Recognition (ASLR)
- Sign Language videos
- Sign Synthesis (3D avatars)
- SiGML, SWML, WebSign, etc.
- Automated Sign Language Recognition (ASLR)
- Sign Transcription
- Gloss (vocal language)
- Phonemic Transcription
- Sign Documentation
- HamNoSys, SignWriting, XML-based, etc.
- Gloss to Text Translation
- Text (vocal language)
- Speech Synthesis
- Manual annotation
- Training data
- Gesture scripting
- Sign Language Annotation
- Optical Glyph Recognition (OGR)
- Novel User Interfaces
- Sign Editors
- NLP & Statistical language analysis
- Sign Querying Functionality
<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Conference</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monteiro et al. (2012 SIGACCESS)</td>
<td></td>
<td></td>
<td>Face detection, background subtraction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hand-crafted visual features: velocity-based</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SVM</td>
</tr>
<tr>
<td>Shipman et al. (2015 JCDL, 2017 SIGACCESS)</td>
<td></td>
<td></td>
<td>Face detection, background subtraction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hand-crafted visual features: polar motion profiles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SVM</td>
</tr>
<tr>
<td>Gebre et al. (2013 ICIP)</td>
<td></td>
<td></td>
<td>Face detection, skin detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>hand-crafted visual features</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>random forests</td>
</tr>
<tr>
<td>Yanovich (2016 LREC)</td>
<td></td>
<td></td>
<td>Identification of major sign language constructs: fingerspellings, classifiers, ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hand-crafted visual features</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>k-NN classifier</td>
</tr>
<tr>
<td>Gebre et al. (2014 Comp. Ling.)</td>
<td></td>
<td></td>
<td>Identification of particular sign languages: BSL, DSL, FBSL, FSL, GSL, NGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sparse auto-encoder and 3D CNN</td>
</tr>
</tbody>
</table>
Datasets – the need for sign language detection datasets

• No signing in generic video action recognition datasets, like AVA, THUMOS, ...

• Previous work in SL detection
  • Datasets not made publicly available
  • Small size (~200 videos)

• Sign Language Recognition (ASLR) datasets, Phoenix, SIGNUM, VGG BBC pose, ...
  • Trimmed
  • Captured under constrained conditions

Src: Dreuw et al. (2010)
“Signing in the Wild” dataset

- Untrimmed videos
- Each video can include multiple signing and non-signing events
- Harvested from YouTube

3 categories:
- Signing
- Speaking
- Other

- **1120** video segments
- Each video segment:
  - Up to 6.6 minutes (sampled at 5 Hz)
  - Up to 2000 frames long
- **1.45 million frames** in total

Groundtruthing:
- Frame-level
- 10-frame temporal context
- **1.23 million frames**

Publicly available:
- **IEEE DataPort**
“Signing in the Wild” dataset

Candidate list of 38,000 video URLs

Random sampling of a subset of candidate videos

Automated keyword-based search on YouTube

Manual vetting of candidate videos (inappropriate content, very low-res & poor quality videos filtered out)

Videos truncated to a length of 2000 frames (sampled @ 5 Hz)

List of vetted videos

“Signing in the Wild” dataset

1120 video segments

Groundtruth data

Manual groundtruthing of video frames

<table>
<thead>
<tr>
<th>class</th>
<th>total frames</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>signing</td>
<td>535,105</td>
<td>43%</td>
</tr>
<tr>
<td>speaking</td>
<td>511,446</td>
<td>42%</td>
</tr>
<tr>
<td>other</td>
<td>186,007</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>1,232,558</td>
<td>100%</td>
</tr>
</tbody>
</table>
Example frames from class signing
Example frames from class other
• Automated extraction of features using a Convolutional Neural Network (CNN)

• Combining both visual features and motion features

• Use of a Recurrent Neural Network (RNN) to handle the dynamic temporal patterns present in sign languages
• Two-stream approach (Simonyan, 2014)
Proposed approach

• Motion stream:
  • Performance vs. computational efficiency

• Investigated:
  • Optical Flow
  • Motion History Images (MHI)
  • Multi-frame differencing
**CNN features**

- **CNN streams**
  - Pre-trained VGG-16 (Simonyan 2014)
  - CNN features:
    - ① 7 x 7 x 512 = 25088 feature map from ‘block5_conv3’ layer
    - ② 4096 feature map from ‘fc1’ layer

- **Motion stream CNN features**:
  - We use transfer learning from a distant task (unrelated data) vs. Training from scratch (Yosinski et al., 2014)
  - No fine-tuning of VGG-16 layers
• Optical flow
  • Dense optical flow (Farnebäck’s algorithm)
  • Encoded as RGB
    • Flow vector magnitude → luminance channel
    • Flow vector angle → chrominance channels

Motion data

Motion stream

CNN

Filter weights of first layer replicated

224 x 224 x 30
Motion data

- Multi-frame differencing
- Motion History Images
  - 5 frame temporal window

Motion data

\[
\begin{align*}
&\text{Motion data} \\
&t\quad t+1\quad t+2\quad \ldots\quad t+10 \\
\end{align*}
\]

\[
\begin{align*}
&\text{224 x 224 x 10} \\
\end{align*}
\]

\[
\begin{align*}
&\text{Motion stream CNN} \\
&\text{Filter weights of first layer averaged to 1-channel, then replicated 10 times}
\end{align*}
\]
RNN

- Various RNN options: LSTMs and GRUs
- Stacked RNNs
  - 2-layer GRU
  - 256 hidden units
  - 20 timesteps
    - (2.5 seconds with a 5Hz sampling rate)
RNN training

- Stratified partitioning of the dataset
- Video frames from a single video appear in only one partition
- 5 fold cross-validation

- Mini-batch stochastic gradient descent (SGD)
- Adam optimizer
- Training for 500 epochs, with early stopping (validation cross-entropy loss)

- Training strategy:
  - Initial mini-batch size of 32, learning rate of 0.001
  - Reduce learning rate when validation loss stops improving for the current combination of mini-batch size and learning rate
  - Increase mini-batch size when no more change in validation loss is observed for the given mini-batch size despite the changes to the learning rate
Results

- Evaluation of different feature maps from the CNN network

<table>
<thead>
<tr>
<th>CNN layer</th>
<th>Feature size</th>
<th>Loss (validation set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-16 block5_conv3</td>
<td>A 25088</td>
<td>0.6681</td>
</tr>
<tr>
<td>VGG-16 fc1</td>
<td>B 4096</td>
<td>0.5037</td>
</tr>
</tbody>
</table>

![Diagram showing different layers and their feature sizes and losses](image_url)
 Results

- Evaluation of the individual performance of the different streams, and when fusing both the motion and RGB streams together

<table>
<thead>
<tr>
<th>Modality</th>
<th>Loss ↓</th>
<th>Accuracy ↑</th>
<th>Time (ms) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB stream only</td>
<td>0.5128</td>
<td>85.01%</td>
<td>–</td>
</tr>
<tr>
<td>optical flow only</td>
<td>0.5387</td>
<td>83.12%</td>
<td>57.8</td>
</tr>
<tr>
<td>MHI only</td>
<td>0.5445</td>
<td>83.67%</td>
<td>17.1</td>
</tr>
<tr>
<td>multi-frame diff. only</td>
<td>0.5738</td>
<td>84.08%</td>
<td>9.7</td>
</tr>
<tr>
<td>RGB stream + optical flow stream</td>
<td>–</td>
<td>87.67%</td>
<td>–</td>
</tr>
<tr>
<td>RGB stream + MHI stream</td>
<td>–</td>
<td>87.60%</td>
<td>–</td>
</tr>
<tr>
<td>RGB stream + frame diff. stream</td>
<td>–</td>
<td>85.61%</td>
<td>–</td>
</tr>
</tbody>
</table>
• Evaluation of different RNN architectures

<table>
<thead>
<tr>
<th>RNN</th>
<th>layers</th>
<th>trainable parameters</th>
<th>Loss (valid. set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>1</td>
<td>4,474,627</td>
<td>0.5144</td>
</tr>
<tr>
<td>LSTM</td>
<td>2</td>
<td>4,999,939</td>
<td>0.6413</td>
</tr>
<tr>
<td>LSTM</td>
<td>3</td>
<td>5,525,251</td>
<td>0.5714</td>
</tr>
<tr>
<td>GRU</td>
<td>1</td>
<td>3,360,259</td>
<td>0.5267</td>
</tr>
<tr>
<td>GRU</td>
<td>2</td>
<td>3,754,243</td>
<td><strong>0.5037</strong></td>
</tr>
<tr>
<td>GRU</td>
<td>3</td>
<td>4,148,227</td>
<td>0.6028</td>
</tr>
</tbody>
</table>
Ablation studies on the proposed RNN network

<table>
<thead>
<tr>
<th>Model settings</th>
<th>Cross-entropy loss on validation set ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>proposed model</td>
<td>0.504</td>
</tr>
<tr>
<td>no batch normalisation</td>
<td>0.609 (≈ 20% increase in loss)</td>
</tr>
<tr>
<td>no dropout layer</td>
<td>0.715 (≈ 42% increase in loss)</td>
</tr>
<tr>
<td>no GRU dropout</td>
<td>0.693 (≈ 38% increase in loss)</td>
</tr>
<tr>
<td>no classifier fc1 layer</td>
<td>0.649 (≈ 29% increase in loss)</td>
</tr>
<tr>
<td>with dropout layer rate: 0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>loss: 0.605</td>
<td>0.577</td>
</tr>
<tr>
<td>with GRU dropout rate: 0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>loss: 0.628</td>
<td>0.601</td>
</tr>
</tbody>
</table>

The results show a decrease in cross-entropy loss when dropout and GRU dropout layers are added to the model, indicating improved performance.
Results

• Comparison with the state-of-the-art in sign language detection
  • ≈ 18% improvement over baseline
  • ≈ 9% improvement when using an RNN versus SVM

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature type &amp; Classifier</th>
<th>Loss ↓</th>
<th>Precision ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline method †</td>
<td>hand-crafted features + SVM</td>
<td>1.114</td>
<td>69.23%</td>
</tr>
<tr>
<td>baseline+RNN</td>
<td>hand-crafted features + RNN</td>
<td>0.841</td>
<td>78.02%</td>
</tr>
<tr>
<td>CNN+SVM</td>
<td>2-stream CNN features + SVM</td>
<td>–</td>
<td>79.15%</td>
</tr>
<tr>
<td>our method</td>
<td>2-stream CNN features + RNN</td>
<td>0.573</td>
<td>87.67%</td>
</tr>
</tbody>
</table>

† Shipman et al. (JCDL 2015, ACM SigAccess 2017)
Results

• Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Actual label</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sign</td>
</tr>
<tr>
<td><strong>Predicted label</strong></td>
<td></td>
</tr>
<tr>
<td>sign</td>
<td>4992</td>
</tr>
<tr>
<td>speak</td>
<td>681</td>
</tr>
<tr>
<td>other</td>
<td>65</td>
</tr>
</tbody>
</table>
Results

Video “YeTuAxh0LZo”

Sample video frames (every 100)

Groundtruth + prediction strips (signing, speaking, other)
Results

Video “VgiM1CFgpbA”
Results

Video “VehhDKHe5Ko”

Second part of this video contains hand motion, clapping and singing
Results

Video “Vbm3MprH3KQ”

Two signers around a table. Several segments mislabeled as speech.
Video “UEfP1OKKz_Q”

Note the boundary errors between speech and other categories.
Results

Video “TWKpeFpbC0w”
Results

Video “wSA84cXmvCA”

A failure case
Results

Video “sWxjJaRj1EE”

Correct detection. Signer is wearing a mask.
Results

Video “HUMEcnkvhJU”

Same person first speaks, then starts signing
Results

Video “wCFoJviRw9k”

Video contains a person doing both signing and speaking
Conclusion & Future Work

• A new dataset “Signing in the Wild”
• Successful sign language detection via a two-stream CNN+RNN
  • \( \approx 18\% \) improvement over state-of-the-art
• Visual + Motion features are both important for signing

• Future Work:
  • Signer localisation
  • Identification of particular sign languages & sign language constructs
  • Investigating what the CNN+RNN is basing recognition on
• Thank you for your attention