Unrestricted automatic fingerspelling recognition from video

•••

Taehwan Kim, Jonathan Keane, Weiran Wang, Hao Tang, Greg Shakhnarovich, Jason Riggle, Diane Brentari, and Karen Livescu

This talk

- Project aims
- Data collection
- Data annotation
- Model development
- Next steps

Project aims

Automatic sign (and fingerspelling) recognition

There is relatively little sign recognition work.

- Most work is on constrained vocabulary
- One of the biggest stumbling blocks is handshape recognition.

We restrict ourselves to fingerspelling

- Fingerspelling is (almost) all handshape contrasts
- But, it is constrained in location and movement
- In a lexicon-free environment

Data collection

Our data

Signers

5 signers: 4 are Deaf of Deaf parents, and native users, and 1 is an early learner.

Video

- 2 video cameras recording at 60fps.
- We collected a number of sessions for each signer most at a normal, conversational speed, and some at a careful speed.
- The video was then post processed and compressed for coding.

Our data

Word lists

There were three word lists:

1. A variety of words including English nouns, English names, and non-English words.
2. The 300 most common nouns in the CELEX corpus
3. 300 mostly non-English words, designed to get all bi-grams not seen in the lists above.

Each word was fingerspelled twice in each speed.

Multiple views, multiple signers

Data annotation

The coding process needs to be

Accurate

- Accurate, detailed data is necessary for any linguistic analysis.

Reproducible

- Coding should be able to be reproduced, and individual coders should form some sort of consensus.

Quick

- Coding time is often directly related to the amount of data available to us.

Easy

- A coding system that requires little specialized training is better than one that requires experts to use. (All else being equal)

3-step annotation process

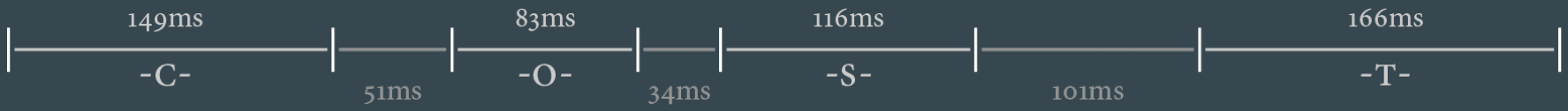
1. Quick pass by three annotators pressing buttons whenever there was a letter
2. Align these quick passes automatically and add in best guesses at letter identity
3. Verify the annotations and mark beginning and end of handshape stability

This process is very reliable. A subset of the data was double coded, and coders agree:

- 61% of annotations have no difference
- Mean difference in times is 2.28msec
- Letter identification has a cohen's κ of 0.9625

Holds and transitions for C-O-S-T



Data annotated

We have 4 signers annotated so far:

- 3,684 word instances
- 600 different words
- 21,453 peaks in total

In addition, there is 3-4x this amount of data not yet coded.

Model development

Hand segmentation

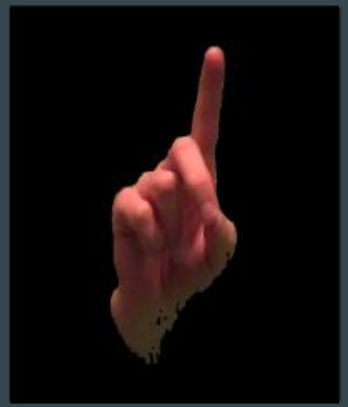
Separate the hand from the rest of the video using color.

1. Manual annotation of 30 frames (per signer)
2. Each pixel is scored for p(hand) or p(not hand)
3. Ignore the face (using Viola-Jones face detector)
4. Ignore when p(hand) > p(not hand) but p(hand) is low
5. Ignore pixels outside of reasonable area of seating

A

L

Y



D

E

F

Handshape descriptors

Resize hand to 128x128 pixels

Computer histograms of gradients (HOG features) on 4x4, 8x8, and 16x16 spatial grids

8 orientation bins per grid cell

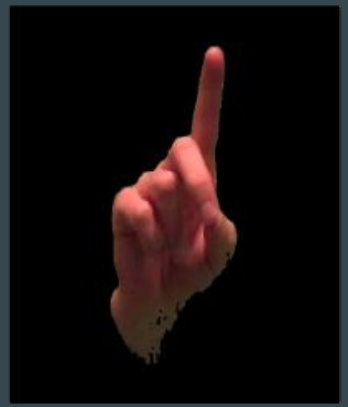
2688-dimensional descriptors

(For speed, HMMs were limited to 200 principal dimensions)

A

L

Y



D

E

F

Models used

- Dynamic Neural Networks (DNNs)
  - Classifying frames
  - Input: image features
  - Output: probabilities of unit labels (or phonological/phonetic features)
- Tandem Hidden Markov Model (HMM) as a baseline
  - trying to identify monograms (not bigrams or trigrams)
- Segmental Conditional Random Fields (SCRFs)
  - Rescoring: rescore the outputs of HMMs
  - First pass method: ignores the outputs of HMMs

Signer dependent recognition

For signer dependent recognition:

- HMMs have 14.6% letter error rate
- Rescored SCRFs have 11.5% letter error rate
- First-pass SCRFs have 8.8% letter error rate

Results: R-O-A-D

GT	<s>	R	O	A	D	</s>
Tandem	<s>			A		</s>
re-SCRF	<s>	U	O	A		</s>
1p-SCRF	<s>	R	O	A	D	</s>

Signer independent models

For signer independent recognition:

- HMMs have 57.2% letter error rate
- Rescored SCRFs have 55.3% letter error rate
- First-pass SCRFs have 60.6% letter error rate

Adaptation of the DNNs

1. Using ground truth data
    - HMMs have 22.0% letter error rate
    - Rescored SCRFs have 21.7% letter error rate
    - First-pass SCRFs have 17.3% letter error rate

2. Word (not segment) labels, run through the signer-independent recognizer, use those as labels, refit.
    - HMMs have 33.6% letter error rate
    - Rescored SCRFs have 32.0% letter error rate
    - First-pass SCRFs have 30.3% letter error rate

Models trained to detect phonological features

Models were also trained to classify phonetic features and the combination of phonological features and letter identity as well as phonetic features and letter identity.

In the signer dependent setting, the letter identity models only were best.

In the signer independent setting phonological or phonetic features are helpful in addition to letter identities for two of the signers.

Conclusions

Using a small corpus of hand-annotated training data, signer-dependent fingerspelling recognition can achieve as low as 8.8% letter error rate.

Signer-independent recognition is much worse, but with a small set of training or enrollment data (as small as 30 words), rates can be brought down to the 17-30% letter error rate range.

Training classifiers on phonological or phonetic features alongside labels improves recognition for signer-independent models modestly.

Next steps

Future work

Gather data from many more signers to see if wide-signer training can help with signer-independent recognition (in progress)

More gold-standard annotation of the data we have already collected

Develop methods of hybrid recognition-human annotations systems

Release the corpus of data and annotations for researchers to use

Thank you!

This work could not be possible without the contributions of our Deaf colleagues: Andy Gabel, Rita Mowl, Drucilla Ronchen, and Robin Shay.

Additionally, the help of many RAs at both University of Chicago and Toyota Technological Institute at Chicago: Susan Post-Rizzo, Erin Dahlgren, Alice Fine, Julia Goldsmith-Pinkham, TJ Heins, Katie Henry, Rachel Hwang, Linda Liu, Rachel Perry, and Katina Vradelis

NSF grants NSF-1433485 and NSF/BCS-1251807