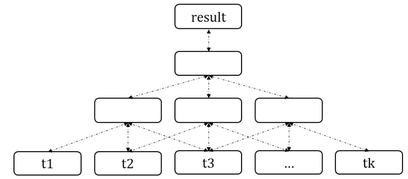
Complexity Beyond the Trigram: Identifying Sign Languages from Video Using Neural Networks

Pamela Toman



Why?

Language identification is a major challenge facing automatic collection of large-scale sign language corpora for machine translation and corpus linguistics work. However, the task has been minimally studied.

What?

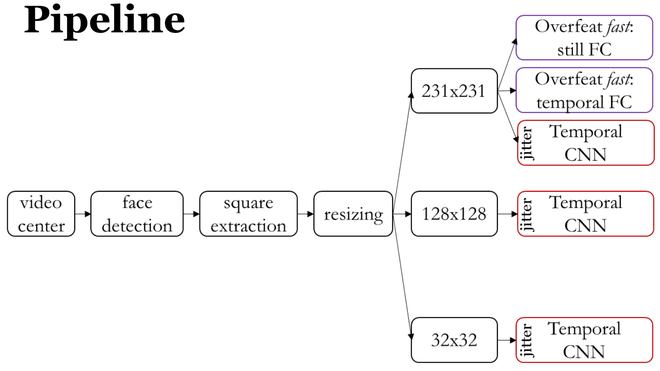
A benchmark dataset is collected and published for three languages: American Sign Language (ASL), British Sign Language (BSL), and German Sign Language (DGS). Evaluations address the effect of temporal information. Preliminary human results substantiate and contextualize the problem.

Data

This project releases SLANG-3k, a benchmark dataset for sign language identification on the internet. It contains 3000 clips of 15 seconds each from attributed Creative Commons videos. Each source contributes only to train, val, or test.

	Clips	Sources	Hours Sampled	Themes
ASL	1000	103	7.7	religion, infomercial, linguistics
BSL	1000	34	9.8	government, medical
DGS	1000	48	6.9	religion, banking, conventions

Pipeline



Experimental Results

The task given a **single frame** is indeed hard.
Accuracy: Overfeat: 54.6% Human: 49.7%

	ASL	BSL	DGS	Unknown
ASL	51.3	5.3	2.3	22
BSL	28.7	12.7	4.7	23
DGS	19	12.5	7.5	32

Human assessment has 49.7% accuracy on still frames (weighted preponderance of "definitely" or "very likely" judgments; N = 24 respondents on 10 items).

Several of the pictures were blurry or the hand shape/position was unclear, making the task difficult, if not impossible.

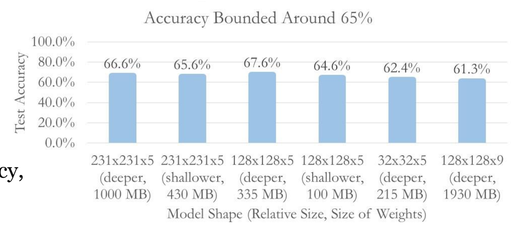
I have used ASL for 23 years. I know that while some signs may be similar to other sign languages, it is very hard to know the sign from a still photograph, even if it is in the language in which you are fluent.

Including temporal information improves performance. Six neural networks were trained.

There seems to be a bound around 65% accuracy when using 5 frames of context. Since 5 frames is still too few to identify most morphemes, performance is expected to increase with additional temporal information.

	ASL	BSL	DGS
ASL	133	15	59
BSL	17	184	60
DGS	62	5	137

Neural network 128x128x5 (deeper) has 67.6% accuracy, compared to 54.6% on Overfeat-based stills.



How much are results affected by recurring signers or series?
All 5 frame models achieved 100% accuracy on a smaller corpus of 92 ASL clips derived from the NCSLGR corpus, an unrelated and non-overlapping source.

Contributions

- Establishes that temporal information more than offsets less spatial information and less computational power
- Introduces a video classification task for which the presence of temporal information can substantially improve accuracy
- Produces SLANG-3k, a public benchmark dataset of sign language identification, and releases baseline scores for future work

All results and images are from the 128x128x5 (deeper) model unless otherwise specified.

Qualitative Evaluation

For many images, linguistically meaningful regions are important to the final score.

Green regions are important to correct prediction.

Clear regions are important to differentiation.

The signer and his BSL-indicative two-handed fingerspelling matter.

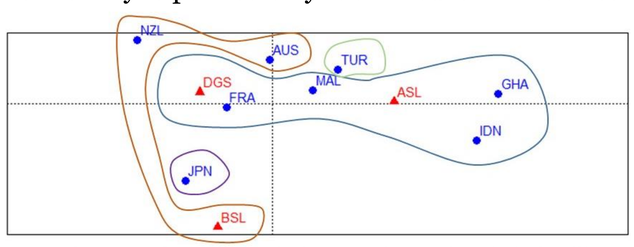
For videos in a series, the network sometimes finds the background more informative than the language.

Green regions are important to correct prediction.

Clear regions are important to differentiation.

The signer is unimportant compared to his infomercial background.

The ASL family is particularly identifiable in mistakes:



Correspondence analysis on confusion matrices. Linguistic families:
ASL, *DGS*, FRA, MAL, GHA, IDN
BSL, AUS, NZL TUR JPN

Future Work

- Train deeper and more effective networks with better hardware
- Reduce the tendency to rely on background information through collecting a larger, more diverse dataset and/or resuming work on background removal