# SignNet: End-to-End ASL to Text Translation

Adith Balamurugan, UC Berkeley

# OVERVIEW

## Objective

Sign language is an important means for communication between the deaf community and the majority speaking population. This project presents a method for automated **American Sign Language (ASL)** to text conversion. There are existing methods which currently combat this problem using deep learning, however, nearly all of these methods focus on the recognition and translation of single letters and digits (aka fingerspelling). In this project, I use a deep learning architecture to take variable length raw video input and output a text translation. This translation is not restricted to solely letters and digits, but generalizes to more complex words and phrases commonly used in conversation. My method draws important features from each frame of the video and is able to detect and translate many word signs as well as numbers and English letters. A few example words include “you”, “eat”, and “work.”

## American Sign Language

Every ASL word/sign is comprised of **5 key elements:**

1. **Handshape**
2. **Movement** (direction of motion)
3. **Location** (in my case position of hand in frame)
4. **Palm Orientation**
5. Non-manual marker (e.g. **facial expression**, shoulder tilt)

## Contributions

- Translation beyond fingerspelling for single letter or digit classification
- Translation for word/phrase level gestures
  - **Single hand** gestures
  - **Symmetric two hand** gestures
- Apply methods from recent lip-reading research to ASL translation objective

# APPROACH

## Inspiration

The approach used in this paper is strongly motivated by the work done in the **LipNet** paper, which focuses on lipreading from video and converting to text. I borrow a **frame level classification** to video level translation method used in this (and other similar) works.

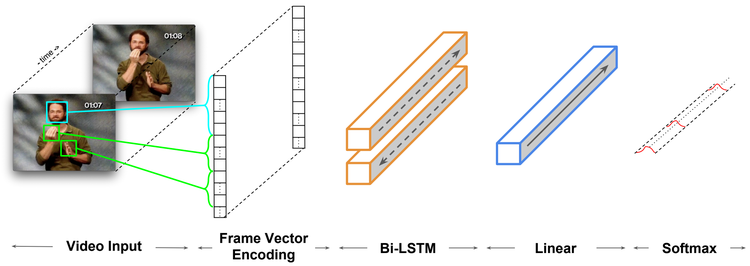
## Data

In order to train and evaluate the method, I took advantage of the **American Sign Language Lexicon Video Dataset (ASLLVD)**
The ASLLVD consists of videos of 3300+ ASL signs, each performed by 1-6 ASL signers, each taken from up to 4 different angles. The videos are annotated with **unique gloss for each sign variant**, lexical variants, indication of repetition, handshape (for dominant and nondominant), and timestamps for videos containing more than 1 sign

| Class of signs | Number of signs | Number of sign variants | # sign variants with { 1, 2, 3, 4... } consultants | | # tokens (examples) per sign { 1,2,...,6, >6 } | | Number of sign tokens | |
|---|---|---|---|---|---|---|---|---|
| Monomorphemic lexical signs | 2,284 | 2,793 | x1 | 621 | 587 | x1 | 8,585 | Two-handed |
| | | | x2 | 989 | 858 | x2 | | 5,713 |
| | | | x3 | 394 | 386 | x3 | | 67% |
| | | | x4 | 563 | 491 | x4 | | |
| | | | x5 | 85 | 142 | x5 | | One -handed |
| | | | x6 | 141 | 154 | x6 | | 2,873 |
| | | | | | 175 | >6 | | 33% |
| Compound signs | 289 | 329 | x1 | 129 | 117 | x1 | 749 | |
| | | | x2 | 106 | 107 | x2 | | |
| | | | x3 | 48 | 46 | x3 | | |
| | | | x4 | 33 | 33 | x4 | | |
| | | | x5 | 4 | 11 | x5 | | |
| | | | x6 | 9 | 13 | x6 | | |
| | | | | | 2 | >6 | | |
| Number signs | 76 | 88 | | | | | 260 | |
| Loan signs | 46 | 52 | | | | | 136 | |
| Classifier constructions | 27 | 31 | | | | | 38 | |
| Fingerspelled signs | 21 | 21 | | | | | 25 | |
| ALL | 2,742 | 3,314 | -- | -- | -- | -- | 9,794 | |

# MODEL AND EVALUATION

## Model Architecture




1. Raw Video Input
2. Featurize each frame into vector encoding **5 key elements of ASL**
3. Feature vectors passed as Input to **BiLSTM**
4. Linear pass through outputs of LSTM to resolve erroneous/duplicate frames
5. Softmax activation to classify best gesture for each frame (timestep)
6. Combine classifications to output single translation (video level)

## Evaluation Metric

I chose the **WER** metric as measure of performance

- **minimum** number of word/phrase **insertions, substitutions,** and **deletions** required to transform the prediction into the ground truth, divided by the number of words/phrases in the ground truth

| Unseen Speakers | WER |
|---|---|
| Hearing-Impaired Person (avg) | 47.7% |
| SignNet | 63.3% |

## References and Acknowledgement

1. Carol Neidle, Ashwin Thangali, and Stan Sclaroff.2012. Challenges in development of the american sign language lexicon video dataset ( asllvd ) corpus.
2. Carol Neidle and Christian Vogler. 2012. A new web interface to facilitate access to corpora: development of the asllrp data access interface. In In Proceedings of the International Conference on Language Resources and Evaluation.
3. Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016. Lipnet: Sentence-level lipreading. CoRR, abs/1611.01599.
4. Andy Au and Adam Heins. Automated lip reading using delta feature preprocessing and lstms.
5. T. A. Budi Wirayuda, H. A. Adhi, D. H. Kuswanto, and R. N. Dayawati. 2013. Real-time hand-tracking on video image based on palm geometry. In 2013 International Conference of Information and Communication Technology (ICoICT), pages 241–246.