SpellVision

Scot Bishop, Michaela McMahon, Tyler Vogen
Advisor: Dr. Xia
Mechanical Engineering and Electrical Engineering

NORTHERN ILLINOIS UNIVERSITY
College of Engineering and Engineering Technology

Abstract

The present project serves to design and build a computer vision system called SpellVision that recognizes American Sign Language (ASL) fingerspelling in real time and displays text on a screen. The ultimate goal is to develop a technology that makes two-way communication more accessible between the hearing and Deaf communities. The project consists of three phases. First, videos of ASL fingerspelling were captured and individual ASL fingerspelling letters were identified using visual inspection. A neural network using the long short-term memory (LSTM) architecture were then designed and trained to identify fingerspelling using the video dataset. Finally the neural networks were validated by testing its ability in identifying fingerspelling letters from video clips not seen before by the network.

Introduction

- There exists a *divide* between the hearing and Deaf communities that is yet to be *completely bridged*.
- *Gloves*, and sensors that *fit the body* have been tried and tested.

Figure 1: Gloves that are claimed to translate sign language gestures into audio (Left). iPhone 11 Pro camera setup (Right).

- Wearable devices have been looked down upon due to requiring the Deaf user to *adapt* to a hearing world.
- Utilizing a digital *camera* could provide a more *inclusive* and *low-profile* way of communication.

Methods and Materials

- *Dynamic* fingerspelling *dataset* comprised of over 6,000 unique motions.
- Each clip is manually *pre-processed* by identifying *edges* and *cropping* in MATLAB
- Googlenet *feature detection* image autoencoder used to *vectorize image sequences* via the MATLAB deep learning toolbox.
- LSTM *sequence classification* network based on *DeepSign* architecture.
- *Inferred* sign is displayed.

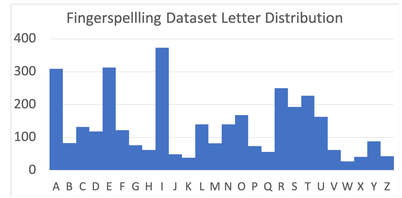


Figure 2: SpellVision ASL Aphabet dataset sign distribution.

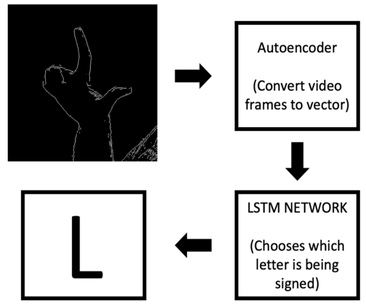


Figure 3: SpellVision data processing pathway. Starting with pre-processed image, to encoding, to classification to display.

Results

- LSTM network was able to *successfully train* on fingerspelling dataset.
- Training stopped at 30 epochs as to avoid overfitting.

```
YPred =

  1×9 categorical array

  N  R  C  G  N  S  T  R  A
  O  R  C  H  E  S  T  R  A
```

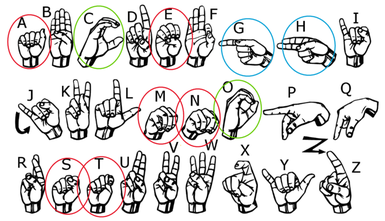
Figure 4: Sample output from SpellVision code; Identifying the word "orchestra". Correct letters shown in red below the computer's guesses.

- The network is able to identify signed letters from *video clips*.
- Some inaccuracy among letters of *similar hand shape.*

Discussion

- Accuracy of about 60%.
- The largest source of error that was observed involved signs that have *similar hand shapes*.

Figure 5: ASL Fingerspelling hand shapes with matching coloured circles denoting signs commonly confused with one another by SpellVision. Notice the similar overall hand shape within the groups.



- Steps to improve accuracy include bolstering dataset with *confused signs* and utilizing a *self-training feature detector* trained on the SpellVision dataset.

Conclusions

- SpellVision can identify *dynamic* ASL fingerspelling in *video* and display the corresponding *text*.
- Inaccuracies exist, but *improvements* can be made.
- Future Direction
  - *Improve detection accuracy*; minimize confusion between letters.
  - Implement base architecture into a *live video feed* application.
  - Analyze *real-time* sign identification response.

Acknowledgements

Dr. Ting Xia, Assistant Professor, Faculty Advisor
German Ibarra, Teaching Assistant
Dr. Robert Sinko, Assistant Professor
All Fingerspelling Dataset Volunteers
The Executive Board of the NIU Deaf Pride Club