Group 7 Final Project Proposal

Machine Learning II, Monday Section
Fall 2018
The George Washington University
October 31, 2018

Bill Grieser
Darshan Kasat
Shivam Thassu

The Question

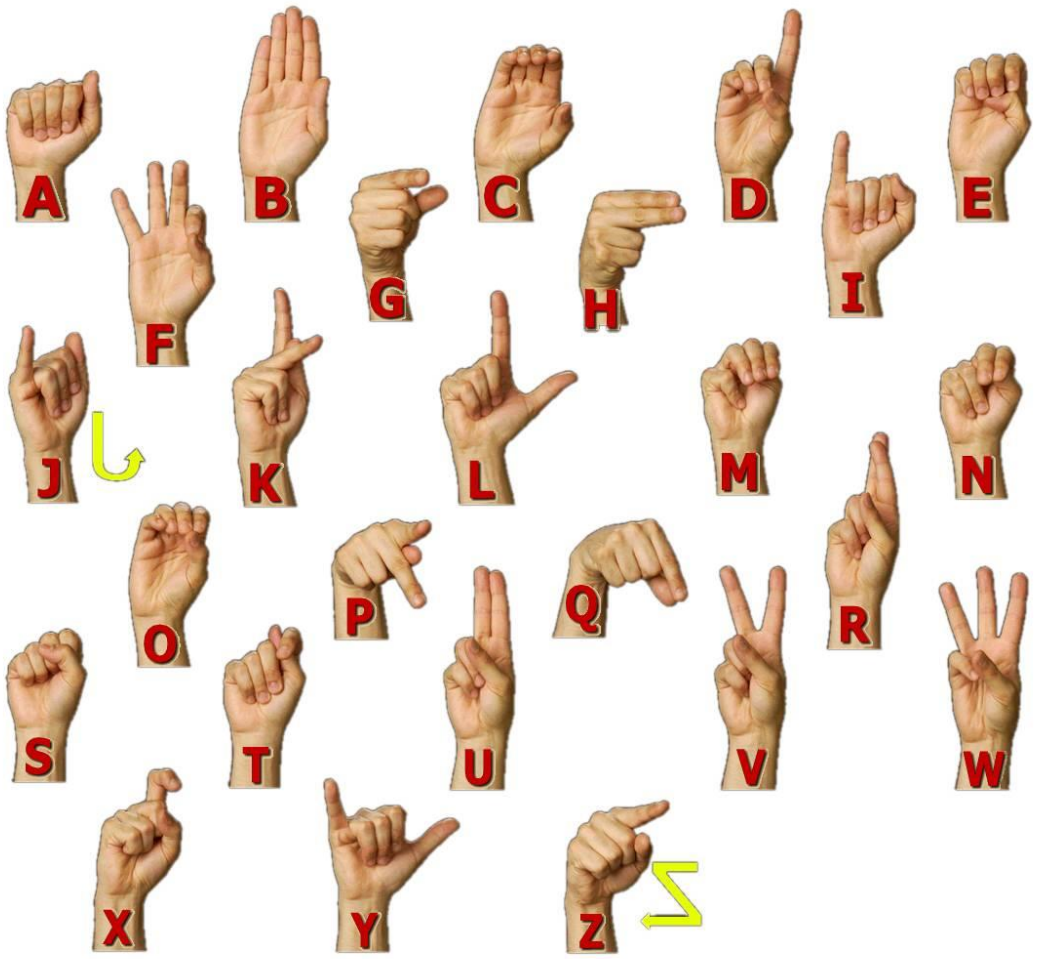
Can we train a model to correctly recognize and classify American Sign Language Alphabet gestures made by real people? And furthermore, can we recognize short common words? We create a trained model and use this to power a program that attempts to recognize sign language letters and words captured live through a web cam on a computer.

For the presentation, we will attempt to demonstrate our program by signing to it words as suggested by the class and instructor.

Constraints:

- Static American sign language letter symbols only (this excludes J and Z which include motion, and accented and other non-English characters).
- Common words are the top N-occurring words from a document corpus, such as a news feed.
- When capturing live images, we may use a "photo booth" approach where the program instructs the signer to prepare a sign and then captures a static image if recognizing gestures from live video is not practical.

This question was interesting to us because it is relevant in many fields, such as gaming and accessibility. One of our team members is a gamer and another is anticipating aging issues, and being able to communicate with computers without using a device is relevant to us. We also like this question because it has elements of both computer vision and natural language processing.



Approach

We will develop multiple neural networks. Our current thoughts on the kinds of network we will use (subject to revision pending further development of the project):

- The network that recognizes gestures will likely be a Convolutional Neural Network that processes images.
- The network that recognizes words will use some kind of delay, such as LSTM.

We have not selected a final framework yet; we are considering Pytorch and Caffe. We will use OpenCV for image manipulation.

We will use Cross-entropy loss to measure model performance.

Data

A source for training data is from Nichoas Pugeault (Pugeault & Bowden, 2011) for a similar project. http://empslocal.ex.ac.uk/people/staff/np331/index.php?section=FingerSpellingDataset

This data includes depth information from the camera used in Nicholas's project. We will determine if we can capture depth data for our live demonstration and this may determine whether we include the depth data.

We may need to generate our own training data through image capture.

Bibliography

Octavio, A., Ploger, P. G., & Valdenegro, M. (2017, October 23). *Real-time Convolutional Neural Networks for Emotion and Gender Classification.* Retrieved October 31, 2018, from github.com: https://github.com/oarriaga/face_classification/blob/master/report.pdf

Pugeault, N., & Bowden, R. (2011). Spelling It Out: Real–Time ASL Fingerspelling Recognition. *Proceedings of the 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision.* Retrieved from http://empslocal.ex.ac.uk/people/staff/np331/publications/PugeaultBowden2011b.pdf

Schedule

Milestone	Date
Detailed Project Plan	November 6, 2018
Data Obtained, cleaned	November 13, 2018
First pass at networks	November 20, 2018
Refined Netoworks	November 27, 2018
Minimum Viable Product -> Input to fit & finish & testing	December 4, 2018
Final Presentation	December 10, 2018