SpellVision 2.0

N. Emerson, E. Taylor, C. Maldonado

College of Engineering & Engineering Technology

Northern Illinois University

Dekalb, IL

***Abstract*—This project was created intending to use computer vision to be able to recognize the American Sign Language (ASL) alphabet in real time with high accuracy. The reason for such a project is to help diminish the gap between those who can hear well and those hard of hearing or even deaf. This can be overcome by creating a large dataset of images that correspond to the letters of the ASL alphabet applied to deep neural networks. These images are labeled according to the letter being signed. They are processed through a neural network using transfer learning to help the machine "learn" what is being signed after already having been taught on larger datasets of many more images and classifications.**

Keywords-American Sign language, neural network, transfer learning

I. Introduction

There is a growing need for a reliable way to communicate with others that have hearing disabilities. There are approximately eleven million people, about 3.6% of the United States' population, that consider themselves to be deaf or have significant hearing issues [1]. There is an organization called the Registry of Interpreters for the Deaf (RID) that can be helpful to those who are in the non-hearing community. Although, there are only about 15,000 recorded members in this organization [2]. Having a means to communicate more effectively where the non-hearing and hearing communities interface could be life changing for those who cannot experience seamless communication with most of the world right now. It is quite the hefty task to create a real-time communication scheme between the hearing and non-hearing communities and this design hopes to be one large step toward that end goal.

II. Methods and materials

Other designs have been attempted to solve the issue of two-way communication between the hearing and non-hearing. Most designs include hardware that can be difficult to carry around or more than one camera. These designs can be useful but are not ideal for those who are hard of hearing to use in many public spaces. Also, these designs limit accessibility by requiring unnecessary hardware. This project creates a simpler implementation that can be used in almost any situation. The hardware that is needed for this design is a camera and a simple cotton glove. The glove is something very simple that can be carried around and not make the individuals using this design standout. Many pictures need to be taken to train a neural network. The pictures that are taken need to represent each of the ASL letters. The letters "J" and "Z" are not trained in the network due to the letters themselves requiring dynamic movements. Transfer learning is used, and the network being used only accepts static pictures. Once the images are taken, a file needs to be created that can describe to the network the letter being signed - as well as where in the image the glove is. These images and extra files are separated into training and testing folders. The testing partition gives a good representation of how the network will identify new images that the network has never seen. So, the training and testing datasets are passed through a previously created neural network using transfer learning. Training of the network can take a full day without a dedicated graphics processing unit (GPU). Once the network is trained, detections can be made in real time.

A. Overall Design

SpellVision 2 takes the live video input from a webcam and scans the frames to detect the hand of a person wearing a color-coded glove. The video frames are processed through a mobilenet single-shot-detector (SSD), and when the desired confidence of a hand sign is met the computer places a box around the hand and states the letter being signed. Along with the SSD box and labeling of the letter being signed, it will also display the confidence it has in its choice. SpellVision 2 achieves this architecture using machine learning and transfer learning.

B. Programs/Software

- Anaconda
- Python 3
- Jupyter notebook
- TensorFlow

C. Components/Devices

- HD Webcam
- Computer
- Latex/cotton glove
- Paint

D. Dataset /Labeling

A large dataset is created that represents the 24 static letters in the ASL alphabet. Each of the letters represents a class to the network. The only letters that aren't represented by this dataset are "J" and "Z" since they include dynamic motion. The networks that are created take about 1,440 images. That means that there are 60 images of each letter created. All the images must be labeled according to the letter being signed. The location of the colored glove and the label of the letter are created and exported to a file with these details being specified in a way the network can read. Each image and label go through a random image augmentation to expand the dataset and create images that seem new to the network being trained.

E. *Training*

Training is done in python using the Machine Learning platform called TensorFlow. The datasets and label files need to be converted into a format that TensorFlow can process to use transfer learning. The folders representing the training and testing data are transferred to tfrecord files. The models that have been created for object detection need to be downloaded to start training a network using transfer learning. The configuration files need to be adjusted slightly to represent the number of classes in the dataset. The classes are each of the letters in the ASL alphabet. So, the configuration file number of classes is 24. The number of training steps used is 20,000 to create high accuracy, but it comes at a cost to training time. The training of a network with 20,000 steps can take around 24 hours without a GPU.

III. Discussion and Results

SpellVision 2, was able to detect someone performing American sign language, with a color-coded glove in real time. To check the performance of the program the performance metrics were obtained, also known as the mAP (mean average precision). Through the performance evaluation it was shown that the ASL translator had a precision of 88.4%.

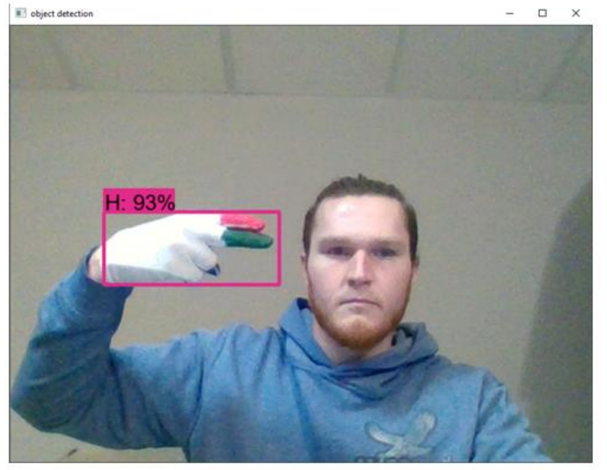
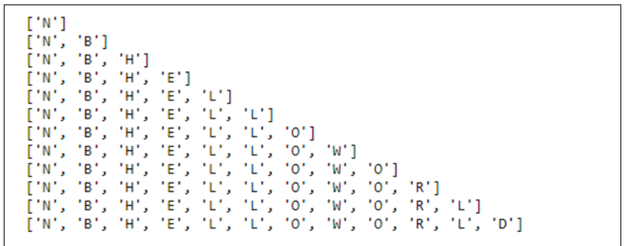


Figure 4. Showcase of live detection with confidence level shown



```
['N']
['N', 'B']
['N', 'B', 'H']
['N', 'B', 'H', 'E']
['N', 'B', 'H', 'E', 'L']
['N', 'B', 'H', 'E', 'L', 'L']
['N', 'B', 'H', 'E', 'L', 'L', 'O']
['N', 'B', 'H', 'E', 'L', 'L', 'O', 'W']
['N', 'B', 'H', 'E', 'L', 'L', 'O', 'W', 'O']
['N', 'B', 'H', 'E', 'L', 'L', 'O', 'W', 'O', 'R']
['N', 'B', 'H', 'E', 'L', 'L', 'O', 'W', 'O', 'R', 'L']
['N', 'B', 'H', 'E', 'L', 'L', 'O', 'W', 'O', 'R', 'L', 'D']
```

Figure 5. Demonstration of letters being saved, to spell out word.

In Figure 3, the chart and corresponding data is given for the mAP score. The ASL translator obtained a score of 88.4% with 20,000 steps. The highest possible score for this test would be 100%, meaning that the translator is correct every single time. Most of the time the network can detect what letter is being signed. There are a few minor examples in which rotating the hand by about 20 degrees will show detections of the wrong letter, but these can be learned easily with more data in the future. Increasing the dataset with well labeled images will increase the accuracy of the system. Many people sign letters a little differently, and thus creating this larger dataset with more people will allow for a little variety - an even better network for the masses.

IV. Conclusion

SpellVision 2's goal was to improve upon the initial SpellVision 1 team's design. This task was accomplished as SpellVision 2 is able to perform live video detection and translation of people using ASL with high accuracy.

A future group for SpellVision will hopefully take it one step further and create a mobile application made available for download on Android and Apple platforms that can classify entire word symbols by integrating facial expressions and relative motion of the hand from the face.

Acknowledgment

The authors would like to thank Dr. Ting Xia with the Department of Mechanical Engineering at Northern Illinois University. We would also like to thank Ian Gilmour for being our teaching assistant throughout the entire process.

References

[1] [1] RIT Libraries. (2011). Deaf Demographics and Employment: Demographics Statistics. Retrieved from https://infoguides.rit.edu/c.php?g=380750&p=2706325

[2] Mitchell, R. E. (2006). How many deaf people are there in the United States? Estimates from the Survey of Income and Program Participation. PubMed, 11(1):112-9. doi: 10.1093/deafed/ENJ004

[3] Renotte , Nicholas, director. *Real Time Face Mask Detection with Tensorflow and Python | Custom Object Detection w/ MobileNet SSD*. *YouTube*, YouTube, 1 Nov. 2020, www.youtube.com/watch?v=IOI0o3Cxv9Q.