SpellVision: A Computer Vision System for the Translation of American Sign Language Fingerspelling to Text

Scot Bishop, Michaela McMahon, Tyler Vogen
Department of Mechanical Engineering
Northern Illinois University
DeKalb, IL United States

***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 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.**

Keywords- American Sign Language; LSTM; Neural Network; sign recognition; feature detector

I. INTRODUCTION

Today, there are an estimated half a million people in the US who use American Sign Language (ASL) as their primary communication method [1] while there are only around 1400 members of the Registry of Interpreters of the Deaf (RID) [2]. There is a clear need for ASL translation to be more readily available so that members of the Deaf Community can have the same opportunities as and convenience as others.

ASL is a full and complete language with its own vocabulary, grammatical structure and culture. The complexity of the language and the people who use it need to be respected. As this project seeks to begin creating a bridge connecting the hearing and Deaf communities it is crucial to remember that full communication is two way. The technology created should not neglect the desires and needs of either party. This bridge should be built meeting in the middle making sure that accommodation is given equally.

The vision for a full, real time, two-way communication platform between signer and speaker is no small task. The intended product of this project, SpellVision is to be the first step towards this larger goal. It set out to create a prototype that can recognize and convert live fingerspelling into text.

II. DESIGN APPROACH

This project takes on a new design approach to the problem at hand. Traditionally, sign language translation devices have required Deaf users to wear or carry excessive hardware. As outlined above, this causes ethical issues by subjecting solely a Deaf user to equip themselves extra hardware in order to accommodate a hearing person [3]. This project adopts a computer vision-based approach utilizing the power and versatility of neural networks for the recognition and conversion of ASL fingerspelling in video to corresponding English letters.

The present product, called SpellVision, is intended for faster, more nuanced fingerspelling like in live conversations. Its design approach adopted the architecture of DeepSign[4], a computer vision algorithm that was designed to recognize full body signing by using a combination of edge detection in images, a self-training image feature encoder and an long short-term memory (LSTM) neural network. Other approaches look at the static form of each letter, whereas DeepSign uses the dynamics of each sign to help predict each letter. Therefore, the LSTM neural network in SpellVision will be trained using many short clips of letters as they form into their recognized shape.

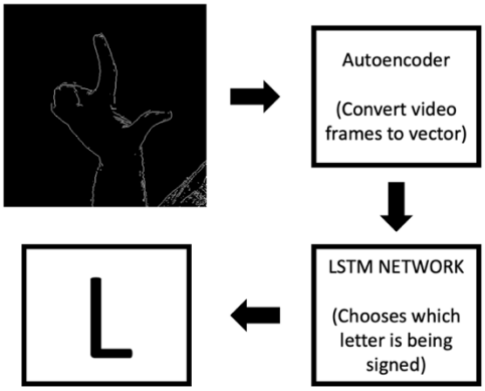

Figure 1: Data through the algorithm to obtain a text output.

As shown in Fig. 1, the process starts off capturing a video clip of a person signing a letter. The video clip can range from only a few frames to more than 40 frames. Edge detection is performed on the clip and is then passed to an encoder. The encoder vectorizes each image in the clip, and then passes those vectors to the LSTM sign identification network. Once a sign is identified, it is displayed on the screen as text.

III. FEATURES

A. A New Kind of Fingerspelling Dataset

To train the neural networks in SpellVision, a dataset consisting of short video clips has been created by visually inspecting and cutting video of volunteers fingerspelling a set

of words. The SpellVision dataset consists of over 3,000 uniquely dynamic signs of the 26 letters of the ASL fingerspelling alphabet (Fig. 2). To account for both right and left-handed individuals, each clip is also flipped horizontally, for a total of over 6,000 categorized letter samples.

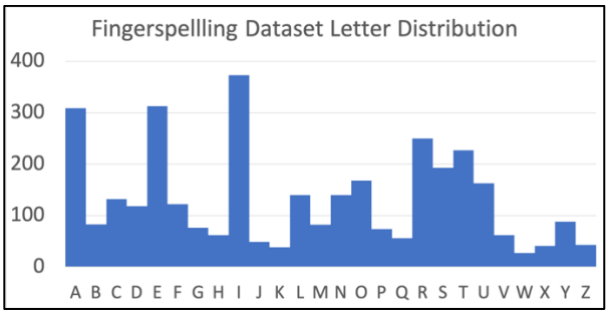

Figure 2: Distribution of letter clips in dataset

In order to simplify the data, edge detection was performed on each video clip before insertion into the network for training. This gets rid of unnecessary components such as color shading and allows the network to focus on the movement of the shape alone over time.

B. A Feature Detector

Because the Long Short-Term Memory network needs a vector of data representing each video clip as an input, a feature detector is needed to compose each frame of the video to a vector, while retaining the important parts of the video. A sequence is then made up with each video frame's vector and can then be passed into the LSTM as one object for each full video clip. The feature detector locates all the important features within each frame of the video for the LSTM to then use to classify each video as it is read in.

C. A Memory Network for Classification

The final step in the process is passing the encoded signs through the LSTM network for classification. The network is trained with a portion of the dataset and then tested with the remainder of the dataset. At each step in the training and testing process, the network verifies the accuracy of each sign and its corresponding label. After training and testing of the LSTM are completed, any new signs are encoded and then passed through the LSTM which outputs an associated letter and probability of that letter being correct.

D. Results

The LSTM network was able to successfully train on the SpellVision video dataset. Training was stopped at 30 epochs to avoid overfitting. When passed a video clip of a single letter being signed, SpellVision is able to correctly identify signs with an accuracy of about 60%. Letters that are consistently identified are of more unique shape, while letters that get confused are of similar shape. For example, letters such as J, L, I, R, and W are identified accurately. However, E, M, N, S, and T can be confused with one another, as the hand must form a fist-type shape to form each letter.

IV. CONCLUSION

A. Discussion

It was seen that signs with similar hand shapes were often confused for each other. This may point to a deficiency in the chosen feature detector. The vector fed to the LSTM may not contain the specific information needed for consistent recognition of the confused signs. Despite the low network accuracy, it was seen that the LSTM could recognize signs and like hand shapes. This shows promise for the network with future development of the autoencoder.

The results demonstrate that not only can the network recognize hand shapes, but it can identify signs that are changing with time. The use of an LSTM is a novel approach to fingerspelling, and our design proves this concept.

B. Future Work

There is still a long way to go to reach a full, two-way, real time communication platform. Improvements need to be made in the accuracy of the SpellVision network. Real time hand detection needs to be developed and incorporated. These steps to improve the SpellVision project are still just the beginning to a future with computer vision ASL translation. One component that has yet to be fully developed is the hand detection network [5]. This has been attempted throughout this project; however, results were not good enough to implement into the rest of the algorithm. In addition, the algorithm still needs to be integrated into a device equipped with a camera so that the real-time, real-world performance of the system can be evaluated.

ACKNOWLEDGMENT

The group thanks our faculty mentor, Dr. Ting Xia and TA, German Ibarra for their support and guidance though this project. The group also thanks Dr. Robert Sinko for providing computing hardware for network training, and all the volunteers that helped film our data set. The group also thanks the executive board of the NIU Deaf Pride Club for their advertising support.

REFERENCES

[1] American Sign Language Program @ The University of Iowa (Department of Speech Pathology and Audiology, 2004) ASLTA (NC ASLTA and NCAD Ad Hoc Committee, 2004) Colorado Department of Human Services (Colorado Commission for the Deaf and Hard of Hearing, n.d.)

[2] Registry of Interpreters for the Deaf, Inc, "2017 Annual Report Member Services," 1 January 2018. [Online]. Available: https://rid.org/2017-annual-reprot/memberservices/.

[3] "UW undergraduate team wins $10,000 Lemelson-MIT Student Prize for gloves that translate sign language," University of Washington, 12 April 2016. [Online]. Available: https://www.washington.edu/news/2016/04/12/uw-undergraduate-team-wins-10000-lemelson-mit-student-prize-for-gloves-that-translate-sign-language/. [Accessed 17 October 2019].

[4] J. Shah, "DeepSign: A Deep Learning Architecture for Sign-Language-Recognition," The University of Texas at Arlington, 2018.

[5] Tensorflow," GitHub repository, 2017. [Online]. Available: https://github.com/victordibia/handtracking