Sign2Sign - A First Attempt

Titus Weng[1], **SingChun Lee**[1]

[1]Department of Computer Science – Bucknell University
Lewisburg, PA, U.S.A.

{t.weng, singchun.lee}@bucknell.edu

Abstract. *At the core of human engagement is communication. While technological advances enable convenient translation for different language speakers to communicate, millions of Deaf, Mute, and Hard-of-Hearing people still face immense hurdles due to the lack of accessible tools to facilitate direct sign language translation. Our project aims to build a Sign2Sign direct immersive translation tool using WebXR that takes input from any accessible camera and produces output in WebXR-supported platforms. This paper presents the preliminary results of direct translation between ten gestures of American and Chinese Sign Languages for sign language translations in an immersive environment.*

1. Introduction

Communication has been an essential need for people since the beginning of human society. Nowadays, millions within the Deaf, Mute, and Hard-of-Hearing population still face immense hurdles due to a world lacking accessible tools to facilitate sign language translation. Many attempts have been made to mitigate this issue between sign language users and non-users, from sign language gloves that detect hand gestures and translate them into words, to software that translates sentences back into skeletal sign language gestures. For different sign language users to communicate, they need to first translate their sign language into text, then into another spoken language, and then back into the target sign language. Currently, there is no effective software that directly translates among sign languages. To address this problem, we started a project to develop an XR tool, Sign2Sign, to directly translate sign languages immersively. In this paper, we present our first attempt at translating ten gestures between American Sign Language (ASL) and Chinese Sign Language (CSL).

2. Related Work

Sign language translation has been a long-studied problem, from traditional rule-based approaches to contemporary deep learning methods. The primary work in this area focuses on spoken-sign/sign-spoken translation. We refer the readers to a recent systematic review for details [Núñez-Marcos et al. 2023]. Our work follows the publicly available real-time multilingual sign language translator [Moryossef 2023], which follows the standard approach to tokenize input video into glosses, then use deep learning methods to translate glosses into text. The spoken-sign translation is similar, which uses machine learning to acquire glosses from sentences and then generate animated avatars from the glosses. This approach requires a text-to-text translation between two spoken languages. Our project aims to merge this three-step approach into a direct translation system. We begin with experimenting with translating ten gestures between ASL and CSL. Our results later enable an immersive XR sign language translator.

3. Method

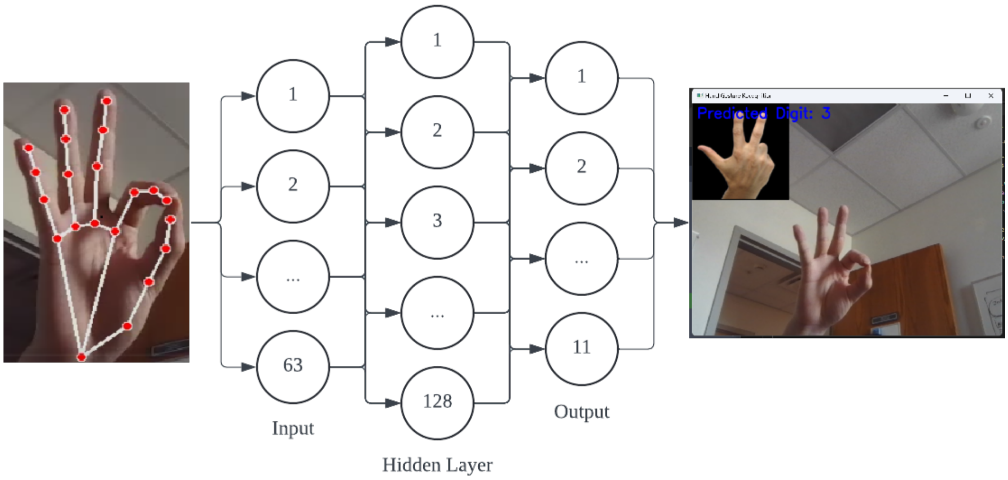


Figure 1. A simple neural network translates glosses detected using MediaPipe (left) to a target sign language (right)

In this first attempt, we focused on translating ten ASL and CSL gestures that represent numbers from 1 to 10. Our system first processes the input video using MeidaPipe [Lugaresi et al. 2019], which detects hand landmarks and returns 21 3D key-points of each hand (Fig. 1 left). A trained simple neural network consisted of a hidden linear layer of size 128 and ReLU activation functions, then takes the key-points and translates them to a probability vector of size 11, which predicts the probability of the ten numbers and a non-digit gesture (Fig. 1 middle). Our software then visualizes the target sign language gesture of the number with the highest probability (Fig. 1 right). If it is a non-digit, nothing is displayed.

4. Dataset and Training

We used PyTorch to implement the neural network [Paszke et al. 2019]. To train the network, we collected from the Internet and created ASL and CSL videos consisting of numbers, and we pre-processed these videos using MediaPipe [Lugaresi et al. 2019], resulting in a total of 15566 ASL and 18219 CSL glosses. We manually labeled these glosses with the corresponding ground-truth vectors. For examples, digit 1 has the ground-truth vector $[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$, for digit 10, it is $[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]$, and for non-digit, it is $[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]$. We randomly selected 80% of the prepared dataset for training with 1000 epochs. The training loss (in blue) plotted in Fig. 2 shows that the network is converged. The prepared dataset will be made public for academic sharing and idea exchange after the publication.

5. Results

To evaluate our system, we used the remaining 20% of the dataset. The accuracy (in red) per each epoch is plotted in Fig. 2. After 1000 epochs, the ASL model obtains an accuracy of 96.08% while the CSL model reaches 91.76%, despite using only a simple neural network. We observed that, even though the training loss is consistently decreasing, the testing accuracy contains spikes, indicating that the training dataset might not be enough

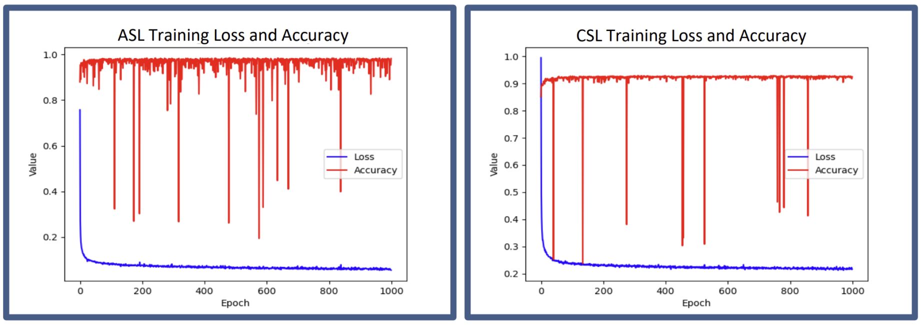
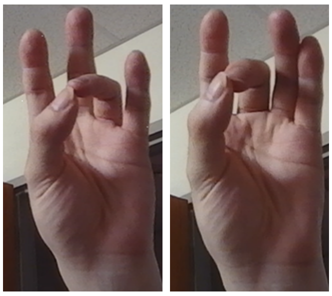
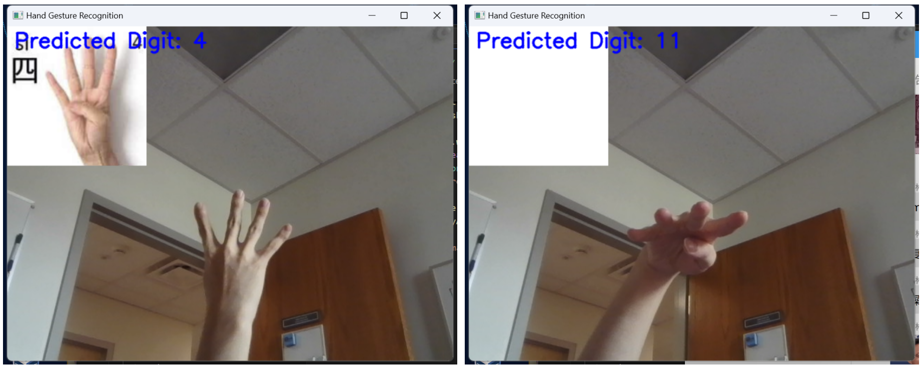


Figure 2. This figure represents the training loss and testing accuracy for ASL and CSL models over 1000 epochs



to generalize a stable model. Occasionally, the model falls into a local minimum that produces worse testing accuracy. The situation is more severe for ASL than CSL. This might be due to the high similarity between numbers 7 and 8, as shown in the inset, which creates more local minimum profiles in the optimization landscape.

Overall, all digits obtain accuracy over 90%; however, CSL model performs worse than ASL. Some digits such as the 10 in CSL presented more challenges because when we trained we found that some poses of the same number were different – even within CSL, users represented number 10 differently. This leads to a lower accuracy rate, and we expect the accuracy to be improved when we use a more sophisticated neural network architecture.



Predicted Digit: 4

Predicted Digit: 11

Figure 3. The system recognizes the gesture successfully in the up-front view (left) but fails to recognize the gesture when viewing from the top-up angle (right).

Fig. 3 demonstrates real-time video Sign2Sign translation. We found that there is still room for improvement. For instance, most of the dataset we used for training was the back of the right hand, so the accuracy was lower than expected when we showed it with the left hand or the front of the hand. In addition, when the hand is placed flat (*i.e.* when the finger is facing the computer), our system fails to recognize the correct gesture. Sometimes, when most of the hand is occluded, MediaPipe fails to determine the hand landmarks. To address this issue, we plan to incorporate more training data in the future.

6. Future Work Towards Direct Sign Language Translator in XR

This paper presented our first attempt to develop a Sign2Sign direct translator using MediaPipe to extract glosses from videos and a simple neural network for translation. While the accuracy is promising, more work is required to build an XR sign language translator.

First, we have not yet achieved our goal of direct translation. In the current attempt, we relied on the fact that digits are universal, so no additional translation was required. In the future, we will explore large language models (LLMs) and redesign our network to train a direct translator from glosses to glosses. We expect more challenges in representing time-series input and output data.

Second, we need more datasets for words and sentences. On top of creating our videos, we will adapt and incorporate publicly available datasets such as Yin *et al.*'s [Yin et al. 2021]. To handle words and sentences, we will need to expand our input glosses from currently one hand (21 key points) to include both hands, the body skeleton, and the face landmarks. Fortunately, MediaPipe supports the detection of both facial and body poses. We plan to keep using MediaPipe as our video-to-gloss pre-processor.

Last, our end goal is to provide an immersive communication experience among Deaf, Mute, and Hard-of-Hearing groups. To promote accessibility, we will develop the gloss-to-avatar control and avatar visualization using WebXR. Our system then supports any VR/AR goggles and mobile devices with a WebXR-supported browser installed. We believe this immersive experience will significantly improve communication among sign language users, providing a seamless and intuitive translation tool similar to how ordinary people use Google Translate.

References

[Lugaresi et al. 2019] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines.

[Moryossef 2023] Moryossef, A. (2023). sign.mt: Real-time multilingual sign language translation application.

[Núñez-Marcos et al. 2023] Núñez-Marcos, A., de Viñaspre, O. P., and Labaka, G. (2023). A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

[Paszke et al. 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

[Yin et al. 2021] Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., and Alikhani, M. (2021). Including signed languages in natural language processing. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.