RECOGNIZING HIGHLY VARIABLE AMERICAN SIGN LANGUAGE IN VIRTUAL REALITY

*Md Shahinur Alam,[1,3], Myles de Bastion[2], Melissa Malzkuhn[3], Lorna C. Quandt[1,3,]**
[1]Educational Neuroscience, Gallaudet University, Washington, DC, USA; {md.shahinur.alam, lorna.quandt}@gallaudet.edu
[2]CymaSpace, Portland, OR, USA; myles@cymaspace.org
[3]VL2 Center, Gallaudet University, Washington, DC, USA; melissa.malzkuhn@gallaudet.edu
* Corresponding author

ABSTRACT

Recognizing signs in virtual reality (VR) is challenging; here, we developed an American Sign Language (ASL) recognition system in a VR environment. We collected a dataset of 2,500 ASL numerical digits (0-10) and 500 instances of the ASL sign for TEA from 10 participants using an Oculus Quest 2. Participants produced ASL signs naturally, resulting in significant variability in location, orientation, duration, and motion trajectory. Additionally, the ten signers in this initial study were diverse in age, sex, ASL proficiency, and hearing status, with most being deaf lifelong ASL users. We report the accuracy results of the recognition model trained on this dataset and highlight three primary contributions of this work: 1) intentionally using highly-variable ASL production, 2) involving deaf ASL signers on the project team, and 3) analyzing the typical confusions of the recognition system.

Index Terms—American Sign Language recognition, virtual reality, gesture recognition

1. INTRODUCTION

Sign languages are natural, full languages developed within the deaf or hard-of-hearing communities. Each signed language uses a set of specific signs and body movements unique to that language. Over 5% (430 million) of the world's population has some form of hearing loss, which is projected to increase to 2.5 billion by 2050 [1]. Signed languages are unique depending on the surrounding culture, ethnicities, and geographical locations where they develop. Most of the world's hearing people are not proficient in signed languages, and thus interpreters are often needed for medical, legal, and educational purposes. As emerging technologies continue to grow, sign language recognition may allow sign language users a more natural way of inputting information into a device. More recently, immersive technologies such as virtual reality (VR) are ripe with educational opportunities, including the potential for learning and interacting with signed languages in VR. Recognition of the users' signing is critical for signed languages to be effectively taught in VR [2]. This paper discusses the recognition of ASL, but our conclusions will also be relevant to other signed languages.

ASL recognition is a growing research field [2]–[9]. Two-dimensional (2D) camera/wearable device-based ASL recognition is the most popular and common approach, yet less efficient and difficult to use in real-life situations [10], [11] because ASL combines hand, face, and body posture with spatial information and dynamic movement. One study found that wearable devices are sometimes troublesome to use [12], and such wearable recognition devices have attracted little interest from signing communities [13]. Overall, a three-dimensional (3D) depth sensor-based camera provides better accuracy and ease [7], [14]. Virtual reality (VR) devices exhibit reasonably good recognition outcomes in some cases [3], [5]. However, none of these are full-fledged VR systems, with many existing efforts primarily dependent on the Leap Motion camera. Hence, standalone ASL recognition in VR remains an unsolved challenge.

Recent research on ASL recognition algorithms typically uses deep learning (DL) algorithms. Survey shows that DL-based algorithms provide superior accuracy [12], [15]. Since VR devices use embedded microprocessors with low computational power, designing a lightweight DL network is crucial. Here, we focused on using a simple network that can be easily computed within a VR environment. As part of our larger work [2], [16], we aim to teach people ASL using a virtual reality game-like environment. In this game, users will enter virtual reality and learn from signing avatars created from motion capture recordings. An example of the 3D environment and the signing avatar is visible in Figure 1.

A critical part of this system is incorporating feedback to inform the users when their sign productions are correct. The feedback relies on capturing and analyzing users' signed productions via the built-in cameras on the VR device. We developed a VR ASL recognition system trained on highly-variable signed input to address previously described limitations. The term "highly variable" indicates that the participants are from different backgrounds, age groups, and levels of ASL fluency. The signs themselves were not inherently "highly-variable"; rather, the production of the signs was not tightly controlled, and they were gathered from a range of signers. Signers were instructed to produce signs

naturally, enabling us to test a recognition algorithm trained on heterogeneous data. Through this study, we aim to provide insights into the relative difficulty and value of gathering ASL data from a small but wide sample of signers.

Figure 1. Avatar teaching TEA sign in a virtual coffee shop.

2. KEY CHALLENGES

2.1. Dataset unavailability

One major problem in the ASL recognition work is that sign datasets are not readily available for ASL, especially in VR environments. Some researchers focus on VR, but the datasets still need to mature [15]. Pugeault et al. published a dataset containing 131,000 ASL alphabet samples collected using the Kinect sensor, OpenNI, and NITE framework [9]. Similar datasets were published by Kapuscinski et al. [5]. Though the datasets are large, they are incomplete. For instance, one dataset includes only 24 characters of the ASL alphabet rather than all 26, and only static signs are included [17]. The ASL alphabet contains both static and dynamic gestures. ASL alphabet signs for J and Z contain dynamic motion; hence they are difficult to recognize with static information alone. In VR, hand gesture recognition is performed from the signer's perspective. These represent significant limitations of the available data.

2.2. Lack of diverse and fluent signers

Most existing ASL datasets for automatic sign language recognition have been collected from hearing participants with low or no proficiency in ASL [20]. New sign language learners typically struggle to produce accurate signs even after years of instruction, with particular errors in movement, location, and orientation of signs [13]. Training a model on signs from novice signers may run the risk of creating homogenous databases, which may contain signs produced in a limited manner--for example, producing the sign in the same location or with the same orientation for every instance of the sign. In the real world, ASL is used by people at many different proficiency levels, with different ways of producing signs, different spatial parameters, and different signing speeds. This natural variability may be one reason why the accuracy of most models' falls in real-life applications. Robust and variable datasets collected from diverse signers are essential for accuracy in practically applied settings.

2.3. Implementation difficulties

Automated ASL recognition ideally involves capturing and computing hand, body, and gaze movements. However, computer vision approaches face several challenges, including occlusion, variable distance from the capturing device, lighting conditions, and color ambiguity. As a result, an ASL recognition system must be robust and able to categorize these nuanced variations in sign production accurately.

Given these existing limitations in the field of ASL recognition, here we trained a VR ASL recognition system using highly-variable signed input. We opted to give signers the instructions to produce signs naturally, with the goal of testing a recognition algorithm trained on heterogeneous data. With this case study, we hope to clarify some of the challenges.

3. METHOD

In this work, we have employed the Oculus Quest 2 as a VR device and MiVRy [20] Unreal Engine plugin for hand detection segmentation, training, and testing on ASL numbers and a single ASL sign. This plugin is lightweight and easy to fit in the VR environment. As sign language or gesture recognition is relatively new research, related frameworks are not widely available. To our knowledge, MiVRy is the most optimal solution aligned with our requirements. This plugin is lightweight and easy to fit in the VR environment. The signers produced signs naturally, with no additional devices beyond the VR headset. We designed a user interface (UI) to navigate different functions and interact with the virtual textbox. The details are discussed in the following subsections.

3.1. Data collection

The UI is shown in Figure 2. Users can create their own dataset by tapping the gesture name text field. Also, the user can modify the gesture duration. Most of the past research projects in this area used a fixed gesture duration (although the duration varies from person to person); keeping this in mind, we gave more flexibility to the user.

When the participant taps the "Record Stroke" button, the system starts tracking hands and joints for the specified gesture duration time. Next, participants need to tap on the "Train" button to store this gesture in the dataset. As soon as the "Train" button clicked, the system saved the gesture to the local storage and trained the network to detect the gesture using the MiVRy plugin.

Figure 2. The UI of the ASL number data collection system in VR.

3.2. ASL recognition

The user interface (UI) during the recognition phase is similar to that used during data collection. Upon selecting the "Detect Gesture" button, the user sees the UI shown in Figure 3, providing users with options to choose different databases and similarity scores. The similarity score represents a threshold value that can be adjusted between 0 and 100, with a higher score indicating a greater confidence in the recognition results. For instance, in Figure 3, a similarity score of 30 is set, indicating that the system will identify ASL numbers only if the recognition confidence score exceeds 30%. Higher values typically correspond to more accurate detection and identification. The DL model trained during the data collection phase is used for recognition, and the system dynamically allocates parameters to optimize performance for the current dataset.

Figure 3. ASL number recognition UI. The detected gesture is shown in the right window with a confidence value.

4. EXPERIMENTAL SETUP

Figure 4 shows the original experimental setup. Participants wore the Oculus Quest 2 headset and signed ASL numbers 0-10 and the ASL sign TEA. The TEA sign is relatively more complicated than numbers and is a two-handed sign prone to occlusion issue. The UI was visible both on the computer monitor and on the Oculus Quest 2 (software version 44.0.0.169.455). The UI was designed and developed using Unreal Engine 4.27 and MiVRy plugin v2.5 for gesture detection. The system ran on a Windows 11 pro-64-bit operating system with 32GB of memory and an Intel Core i9 3.50Ghz clock speed processor.

We designed the system to work with both left-handed and right-handed participants. Figure 3 shows the UI of our experiment. Participants were free to sign with different palm orientations and locations.


Figure 4. Original experimental setup. Participants can see the UI in VR environments, and their view is mirrored on the computer.

4.1. The dataset

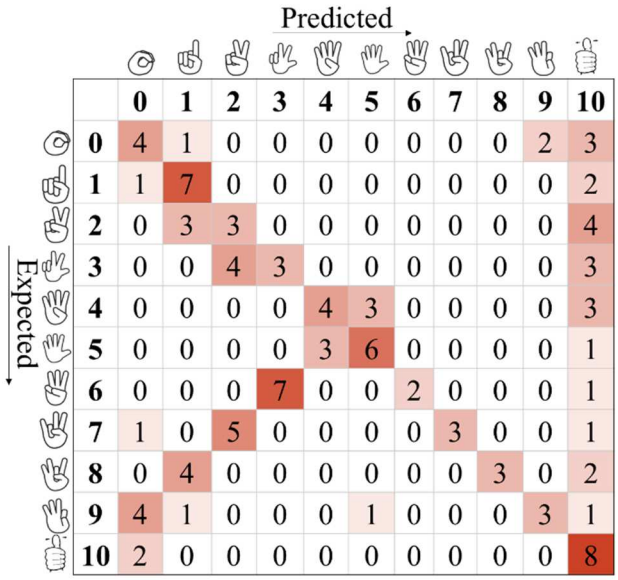
This experiment has two datasets: ASL number signs 0-10 and the ASL sign TEA. The sign duration was fixed (three seconds) for all signs. The number dataset contains 2500 ASL number signs, and the TEA dataset contains 500 signs, collected from 10 participants each. Every participant signed each ASL number (from 0-10, inclusive) 25 times, resulting in 250 signs from each participant. As the TEA is a single sign, participants signed TEA 50 times. Of the ten participants, seven were deaf, one was hard of hearing, and two were hearing. The TEA sign is a complex sign with an occluded hand where both hands are necessary. The purpose of this sign is to verify the model's robustness. We will include more complex signs in our future work.

We recruited ten participants (four men and six women) with diverse backgrounds to train the system on highly variable signed input. The participants, ranging in age from 22 to 46 years, came from various language backgrounds and had between seven months and 43 years of experience signing ASL. Five of the participants had been exposed to ASL since birth. By recruiting participants with varying language backgrounds and levels of ASL experience, we sought to enable the system to recognize a broad range of signing styles.

5. RESULTS AND DISCUSSION

5.1. ASL number dataset

We tested the system ten times for each ASL number and found an average of 46% recognition accuracy; however, the result varies for different numbers. We plotted the results in a confusion matrix in Figure 5 (handshape figures are from the Noun Project created by Stephanie Leeson). We found different accuracy for different numbers with information revealed by the pattern of confusion. The highest and lowest recognition accuracy was found for TEN and SIX, respectively.



Expected \ Predicted	0	1	2	3	4	5	6	7	8	9	10
0	4	1	0	0	0	0	0	0	0	2	3
1	1	7	0	0	0	0	0	0	0	0	2
2	0	3	3	0	0	0	0	0	0	0	4
3	0	0	4	3	0	0	0	0	0	0	3
4	0	0	0	0	4	3	0	0	0	0	3
5	0	0	0	0	3	6	0	0	0	0	1
6	0	0	0	7	0	0	2	0	0	0	1
7	1	0	5	0	0	0	0	3	0	0	1
8	0	4	0	0	0	0	0	0	3	0	2
9	4	1	0	0	0	1	0	0	0	3	1
10	2	0	0	0	0	0	0	0	0	0	8

Figure 5. Confusion matrix of the recognition accuracy. The user's input is plotted in the vertical direction; the horizontal row represents the actual recognized ASL number. The highest and lowest accuracy is found for numbers 10 and 6.

As shown in Figure 5, our results are informative of the typical confusion between similar handshapes. For instance, the ASL sign for SIX uses a handshape with three fingers up, and the recognition model often determines that the signer has produced a THREE (when in fact, they are signing SIX). Similarly, when the signer produced a SEVEN, it was often categorized as a TWO, given that the sign for SEVEN includes the pointer and middle finger raised, just like with a TWO. This pattern extends across several higher number signs; for instance, EIGHT was often recognized as ONE, and NINE was often recognized as ZERO. Thus, the confusion matrix suggests the location of the index finger is over-weighted, whereas the position of the ring and pinkie fingers was under-weighted. The false positive rate can be reduced by improving and implementing more complex DL-based algorithms.

It is essential to note that TEN has the highest recognition accuracy, and it's the only dynamic ASL number sign 0-10. Compared to other research, the proposed one is more accurate for the dynamic gesture [3], [4], [21]. Other researchers focused on image-based recognition, which is better for static gesture recognition. Instead, we focused on trajectory-based recognition. Most signed vocabulary is dynamic in real life; hence, we would expect higher recognition accuracy for other dynamic gestures.

5.2. TEA sign dataset

The TEA sign is a single sign; as a result, we cannot draw a confusion matrix. The average recognition accuracy for TEA was 55%. Occlusion plays a vital role in hand and palm orientation. The occluded finger sometimes incorrectly represents another finger, and it's a challenging task in computer vision research. The accuracy falls when there is an occlusion between fingers; without occlusion, the accuracy goes up to 70%. We expect a better algorithm will provide improved recognition accuracy in future iterations.

6. CONCLUSION

This work has significant practical applications for three main reasons. First, we intentionally captured highly variable sign language productions from a heterogeneous group of ASL users, including variations in hand usage, location, orientation, and movement trajectory. Moreover, this recognition system was trained in virtual reality without the use of specialized cameras or additional devices. This approach more closely approximates the real-world usage of sign-recognition systems, which are less constrained and highly variable. Second, the overall project team includes a majority of deaf members, and all team members know or are learning ASL. This lived experience provides a genuine connection to the language and community. Lastly, the system's typical confusion between ASL digits informs us of systematic patterns in the errors made by the algorithm. We acknowledge that the experimental data is limited to 10 participants, a relatively small sample size. Future work in this area will improve the accuracy of the AI model and allow it to perform better in real-life situations. This work highlights that including the inherent variability of signed language production from the outset is critical for building systems tolerant of real-world variability and leading to better end products. Although initial results may be less "accurate" with more variable input, meaningful progress in this field takes time. Building systems tolerant of real-world variability is critical at all stages. With future work, we aim to train the algorithm and test different signs for use in the VR learning game environment.

ACKNOWLEDGMENT

The authors are grateful to Simone Mangiapelo for assistance with developing the data collection and detection interface. We also gratefully acknowledge the ten participants and team members who gave their time and input to this project. This work is supported by NSF grant 2118742.

REFERENCES

[1] "Deafness and hearing loss." https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed Oct. 02, 2022).

[2] L. Quandt, "Teaching ASL Signs using Signing Avatars and Immersive Learning in Virtual Reality," ASSETS 2020 - 22nd International ACM SIGACCESS Conference on Computers and Accessibility, 2020.

[3] J. Schioppo, Z. Meyer, D. Fabiano, and S. Canavan, "Sign Language Recognition: Learning American Sign Language in a virtual environment," in Conference on Human Factors in Computing Systems - Proceedings, May 2019.

[4] Q. Shao et al., "Teaching American Sign Language in Mixed Reality," Proc ACM Interact Mob Wearable Ubiquitous Technol, vol. 4, no. 4, 2020.

[5] A. Vaitkevičius, M. Taroza, T. Blažauskas, R. Damaševičius, R. Maskeliunas, and M. Woźniak, "Recognition of American Sign Language Gestures in a Virtual Reality Using Leap Motion," Applied Sciences 2019, Vol. 9, Page 445, vol. 9, no. 3, p. 445, Jan. 2019.

[6] T. Starner, J. Weaver, and A. Pentland, "Real-time American Sign Language Recognition using desk and wearable computer based video," IEEE Trans Pattern Anal Mach Intell, vol. 20, no. 12, pp. 1371–1375, 1998.

[7] S. Sharma and K. Kumar, "ASL-3DCNN: American Sign Language Recognition technique using 3-D convolutional neural networks," Multimed Tools Appl, vol. 80, no. 17, pp. 26319–26331, Jul. 2021.

[8] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1114–1119, 2011.

[9] P. Hays, R. Ptucha, and R. Melton, "Mobile device to cloud co-processing of ASL finger spelling to text conversion," in 2013 IEEE Western New York Image Processing Workshop (WNYIPW), pp. 39–43, Nov. 2013.

[10] A. Thongtawee, O. Pinsanoh, and Y. Kitjaidure, "A Novel Feature Extraction for American Sign Language Recognition Using Webcam," in 2018 11th Biomedical Engineering International Conference (BMEiCON), pp. 1–5, Nov. 2018.

[11] F. Wen, Z. Zhang, T. He, and C. Lee, "AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove," Nat Commun, vol. 12, no. 1, Dec. 2021.

[12] P. Barve, N. Mutha, A. Kulkarni, Y. Nigudkar, and Y. Robert, "Application of Deep Learning Techniques on Sign Language Recognition—A Survey," in Lecture Notes on Data Engineering and Communications Technologies, vol. 70, Springer Science and Business Media Deutschland GmbH, pp. 211–227, 2021.

[13] J. Hill, "Do deaf communities actually want sign language gloves?," Nature Electronics 2020 3:9, vol. 3, no. 9, pp. 512–513, Jul. 2020

[14] P. Kumar, R. Saini, S. K. Behera, D. P. Dogra, and P. P. Roy, "Real-time recognition of sign language gestures and air-writing using leap motion," in 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), pp. 157–160, May 2017.

[15] R. Fatmi, S. Rashad, and R. Integlia, "Comparing ANN, SVM, and HMM based Machine Learning Methods for American Sign Language Recognition using Wearable Motion Sensors," 2019 IEEE 9th Annual Computing and Communication Workshop and Conference, CCWC 2019, pp. 290–297, Mar. 2019.

[16] L. C. Quandt, J. Lamberton, C. Leannah, A. Willis, and M. Malzkuhn, "Signing Avatars in a New Dimension: Challenges and Opportunities in Virtual Reality," in Seventh International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual, 2022.

[17] S. N. Reddy Karna, J. S. Kode, S. Nadipalli, and S. Yadav, "American Sign Language Static Gesture Recognition using Deep Learning and Computer Vision," Proceedings - 2nd International Conference on Smart Electronics and Communication, ICOSEC 2021, pp. 1432–1437, 2021.

[18] N. B. Ibrahim, H. H. Zayed, and M. M. Selim, "Advances, Challenges and Opportunities in Continuous Sign Language Recognition," Journal of Engineering and Applied Sciences, vol. 15, no. 5, pp. 1205–1227, Dec. 2019.

[19] D. A. Schlehofer and I. J. Tyler, "Errors in Second Language Learners' Production of Phonological Contrasts in American Sign Language," International Journal of Language and Linguistics, vol. 3, no. 2, pp. 30–38, 2016

[20] "MiVRy 3D Gesture Recognition in Code Plugins - UE Marketplace." https://www.unrealengine.com/marketplace/en-US/product/mivry-3d-gesture-recognition (accessed Oct. 03, 2022).

[21] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using real-sense," in 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), pp. 166–170, Jul. 2015.