Automatic Gloss Dictionary for Sign Language Learners

Chenchen Xu[1,2], **Dongxu Li**[1,2], **Hongdong Li**[1], **Hanna Suominen**[1,3], **Ben Swift**[1]
[1] The Australian National University (ANU) / Canberra, ACT, Australia
[2] Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO) / Canberra, ACT, Australia
[3] University of Turku / Turku, Finland
Firstname.Lastname@anu.edu.au

Abstract

A multi-language dictionary is a fundamental tool for language learning, allowing the learner to look up unfamiliar words. Searching an unrecognized word in the dictionary does not usually require deep knowledge of the target language. However, this is not true for sign language, where gestural elements preclude this type of easy lookup. This paper introduces GlossFinder, an online tool supporting $2,000$ signs to assist language learners in determining the meaning of given signs. Unlike alternative systems of complex inputs, our system requires only that learners imitate the sign in front of a standard webcam. A user study conducted among sign language speakers of varying ability compared our system against existing alternatives and the interviews indicated a clear preference for our new system. This implies that GlossFinder can lower the barrier in sign language learning by addressing the common problem of sign finding and make it accessible to the wider community.

1 Introduction

Unlike most language systems, which are composed of their written and spoken forms, sign languages (e.g., American Sign Language (ASL) and the Australian Auslan language) used by the Deaf or Hard-of-Hearing (DHH) community are represented by the rich inputs including facial and gesture movements. As of the year 2020, 430 million people worldwide have developed hearing loss—that is, one in every ten people—and it is estimated that this number may increase to 700 million by 2050 (WHO, 2021). Sign languages are also used by people suffering the loss of ability to speak (e.g., aphasia) or brain stroke. They are spoken by individuals with various relational connections to sign language speakers, e.g., family members or co-workers. Additionally, a substantial and growing number of people are learning a sign language as a second language, e.g., among U.S. university students (Goldberg et al., 2015). Despite the efforts made in building tools to support their learning (Lee et al., 2005; Schioppo et al., 2019; Hou et al., 2019; Scassellati et al., 2018; Li et al., 2021), many sign language learners have limited means of seeking assistance, and are restricted to class offerings or relying on other experienced sign language speakers. It is therefore increasingly important to support the sign language learner community to facilitate better education and communication.

As a fundamental tool in language study, a dictionary is more than a tool to assist sign language learners in searching unfamiliar words. The rich content present in current online dictionaries (e.g., example pronunciation recordings and visual materials) also provide positive feedback to foster the learner's understanding and proficiency in the target language (Corbeil and Archambault, 2006; Laska, 1993). Most existing sign language dictionaries (e.g., AslSearch (ASLSearch, 2009), Handspeak (Lapiak, 1995), and Signing Savvy (Signing Savvy, 2021)) are text-based and centered on one spoken language, with signs presented in an alphabetical order of their corresponding gloss, i.e., the spoken language counterpart. This does not serve the important scenario when someone encounters an unfamiliar sign and does not know its spoken language translation. Another issue with the text-based dictionary is the fact a one-to-one correlation between sign and spoken language words does not always exist, and no standard convention exists for handling these discrepancies. The absence of these types of dictionary for sign language learner is due to the difficulty of processing visual input and the lack of intuitive alphabetics assumed in most language dictionaries. An early effort made towards a sign-centric dictionary is Tennant et al. (1998) where researchers use pre-defined handshapes (finger poses) to formalise the signs so that they can be arranged similar to a conventional dictionary. Follow-up work (Lapiak, 1995; Neidle

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics System Demonstrations, pages 83 - 92
May 22-27, 2022 ©2022 Association for Computational Linguistics

et al., 2012; Alonzo et al., 2019) parameterised the signs by key properties (e.g., handshape, position of hands, and whether the sign involves repetitive movement) to make a filtering-based search system. In addition, Elliott et al. (2011) used the Microsoft Kinect to collect human body movements from the sign language speaker and match the performed signs against the database.

Modern advancements in deep learning algorithms enable processing of unstructured video inputs, and these algorithms have been applied to sign language. Progress has been made in identifying isolated (Li et al. (2020a,c); Albanie et al. (2020); Sincan and Keles (2020); Momeni et al. (2020)) or continuous signs (Li et al. (2020b); Zhou et al. (2021); Bull et al. (2021); Duarte et al. (2021); Chen et al. (2022)) from a video. This presents opportunity to develop a dictionary system, which accepts direct video inputs from a user performing a sign, and attempts to return the meaning of that sign. One of the early attempts on such video-based system is Alonzo et al. (2019) where the author discussed some characteristics in the design and evaluation metrics regarding the user satisfaction. Notably, the work did not build an actual automatic recognition technique and the users are only presented with a predetermined set of results during the study.

In this paper, we present the platform of *GlossFinder*, our new video-based sign dictionary, where users directly provide videos of the target sign by performing it to their webcam or via uploaded clips, and the system will retrieve matched signs without any extra input. To the best of our knowledge, it is the first attempt of user study with a functioning system built. The study identifies some key considerations in designing for this specific sign language context. It also verifies sign language learners' frustration when using previous sign dictionaries, either due to the steep learning curve or the poor quality of results.

2 System Design

2.1 System Design Criteria

After initial consult with sign language instructors and learners, we determined the following three design criteria items for GlossFinder:

C1. Result with Feedback: Unlike with conventional dictionaries, locating the exact target word (i.e., the sign the user is searching for) is laborious with a sign language dictionary. For example, even basic signs such as those for "father" or "mother" can lead to confusion for beginners. Therefore, the system should provide additional materials to support the user in matching the results and guide refining the search, if appropriate.

C2. Robust to Noise: Example videos in online sign language learning platforms and related research are sourced in a controlled environment. In contrast, we aspire to our system to be applicable to amateur scenarios so that the system is robust even in less formal noisy situations where such clean inputs are unavailable. Namely, we focus on the following two types of noise commonly presented in the videos from informal sources: First, the user captured videos often include blank segments before and after the informative part, in contrast to the videos from professionals which are trimmed and standardized. Second, the lightning conditions may vary among users. We require the system to be robust to these noises.

C3. Minimal Learning: The large sign language learner community includes people of diverse sign language levels and backgrounds. We argue the proposed system should be straightforward to use in order to be perceived as both usable and useful by a broader cohort of sign language learners. With that in mind, a favored dictionary design should be similar to how sign language learners consult peers in practice by performing again the sign to their best ability.

2.2 System Architecture

An overview of the GlossFinder system is demonstrated in Figure 2. Gloss Recognizer accepts an incoming video feed of signers performing the target gloss and determines the ranking of predicted gloss categories. Top gloss candidates from the ranking result are relayed to the *Gloss Retriever* component to collect enriched information for each gloss, e.g., the sample gloss video. Users' access to the system through the web-based platform as illustrated in Figure 1 where they can provide the gloss video feed either by using their camera to record or uploading a pre-recorded video file.

2.3 Gloss Recognizer

The Gloss Recognizer is based on models from supervised training on a sign recognition dataset. Recent works in this field can be categorized into two streams depending on the input feature, namely, human pose or gesture based approaches inspired by the long-term development in sign language rep-

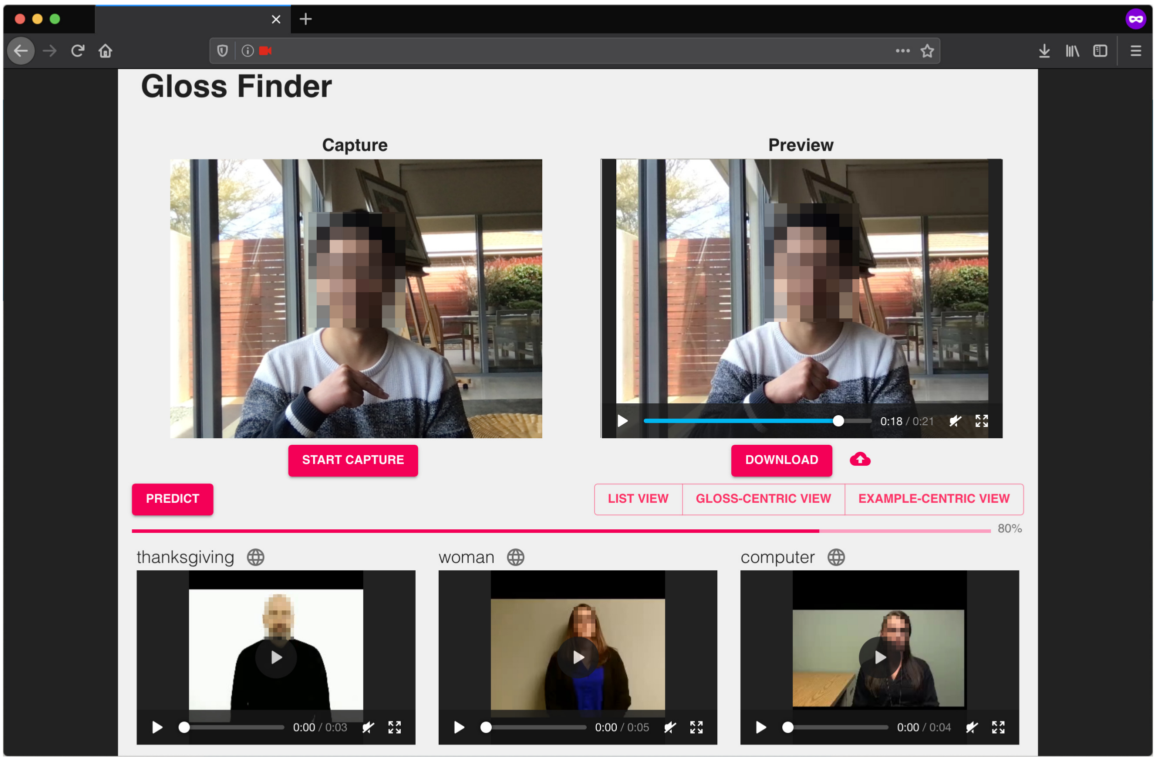


Figure 1: A screenshot of GlossFinder. The top panel is for the input, with the left part being the real-time capture from a webcam or the user's pre-recorded video. The captured video is played to the right as a preview. After the system has made a prediction, the candidate glosses are displayed in the result panel at the bottom along (including example videos). (Faces have been blurred in this paper for privacy reasons.)

resentation (Wang et al., 2010), and recent works relying on deep learning and raw video frames (Li et al., 2020a; Momeni et al., 2020). We observe that the recent video-based approaches report a higher top-k accuracy, but in comparison the overall results from pose-based approaches are more consistent in retrieving visually similar examples which may be attributed to the sparsity of their pose input. To align with the aforementioned criteria of **C1. Result with Feedback** which promotes returning more similar signs for the user to examine, we construct a model of each kind and ensemble the results to retain the benefits of both.

Training Material: To facilitate building the recognition model for the model training, we adopt the public available WLASL dataset (Li et al., 2020a) with $2,000$ sign glosses performed by over 100 signers. This dataset ensures an average of 10.5 examples per sign to support sufficient supervision signals for the model training and vocabulary diversity for our user study.

Image-Based Model: We adopt the I3D networks pre-trained on the Kinetics dataset, considering their effectiveness on sign langauge recognition (Li et al., 2020c) and translation (Li et al., 2020b). The pre-trained backbone enjoys the robustness to varying video conditions, remedying the second noise covered in the criterion **C2. Robust to Noise**. We attach a projector on the representation features extracted from the I3D backbone network. The model is finetuned on the WLASL dataset, and achieves an accuracy of 60.21% at top-5, slightly better than those baselines reported in Li et al. (2020a).

Pose-Based Model: The pose-based model inherits the setting from (Li et al., 2020a) by first extracting the body and 2D keypoints for each frame applying OpenPose (Cao et al., 2019). Considering that face and lip movements are less reliable in the training corpus WLASL, we only use keypoints of the main upper-body with the both hands. The concatenation of all 2D key point coordinates at each frames forms the input feature, before feeding to the Temporal Graph Convolution Networks (TGCN) (Li et al., 2020a). A complete graph is constructed by connecting all key points present in the input features and the TGCN model is trained by learning to aggregate information over this graph of key points.

Sliding Window Inference: Following the cri-

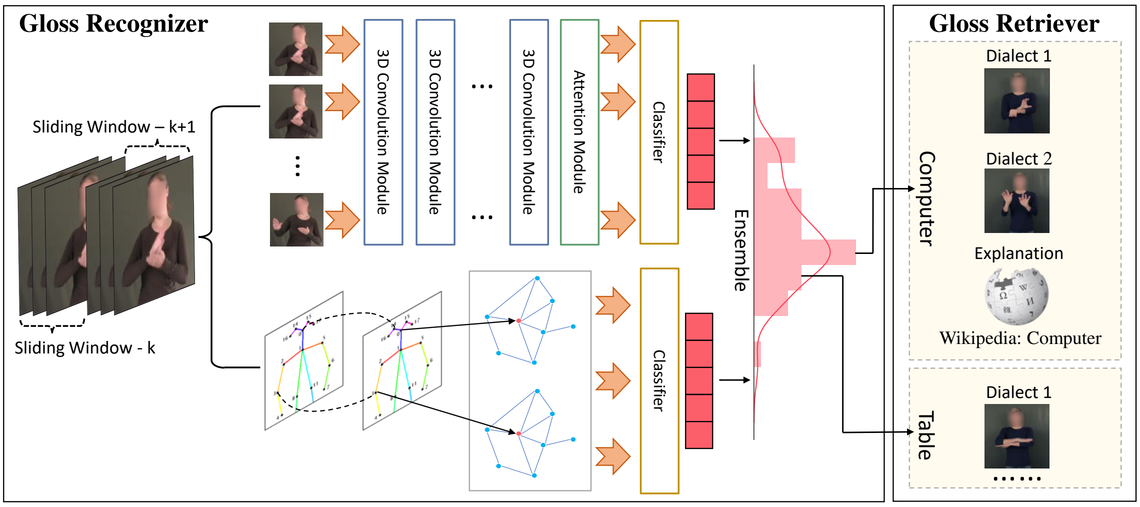


Figure 2: An overview of the main components in GlossFinder. The Gloss Recognizer and Gloss Retriever components jointly generate the enriched view of gloss matches from the input sample video. (Faces have been blurred in this paper for privacy reasons.)

terion of **C2. Robust to Noise**, we incorporate a mechanism to cope with the commonly present short blank segments before and after the representative frames (e.g., users raising or putting down their hands). As demonstrated in Figure 2, We apply the recognition model over several continuous segments of a fixed length of time (sliding windows). As the model slides through the whole input, the result ranking of glosses is obtained from their maximum prediction in all segments. Intuitively, a gloss is predicted as present as long as it appears in any of the segment.

2.4 Gloss Retriever

The *Gloss retriever* component collects enriched information for the predicted glosses from *Gloss Recognizer*, including example videos and explanations. We first shortlist a few public ASL sources and construct a database of glosses with their examples in possible dialects. To avoid duplicate in the result, we only include video examples from the source with most examples present for each individual gloss, with the assumption that these ASL sources mostly include one example video for each dialect. The retriever component includes two result modes. The gloss-centric mode lists individual glosses in the predicted order, with one example video for each. The example-centric mode expands the result gloss with their varieties in the database, that is, a gloss with 3 variety videos will take 3 spots in the result list. The gloss-centric mode provides a clearer view of the gloss guesses from the model, while more freedom is given to the user in example-centric mode to inspect examples and match them with their target in the memory.

2.5 GlossFinder

Based on the criterion of **C3. Minimal Learning**, GlossFinder avoids any pre-defined parameters and the user is only prompted to give a video input of the target gloss, as illusrated in Figure 1. To start, the system guides the user to focus on their camera to capture a video of the target sign. The "capture" button toggles between the start and stop status during the recording. Whenever the stop status is reached, the recorded video is played in the preview window next to it for any adjustment. The preview window is also initialized with an example video to demonstrate the recommended camera position and hand placement. Once the user is satisfied with the recording, they will click the "Predict" button to issue the request to the back-end *Gloss Recognizer* and *Gloss Retriever* component for results. The "capture" button is disabled during the prediction with the progress bar below indicating the current status. Top predicted gloss candidates are then displayed in the result panel sitting in the bottom panel. Each candidate gloss is featured with some example videos from a professional signer so that the users can quickly compare with the one they are looking for.

3 User Evaluation

3.1 Benchmark Systems

We include in our comparison the existing public available sign language dictionaries that can serve

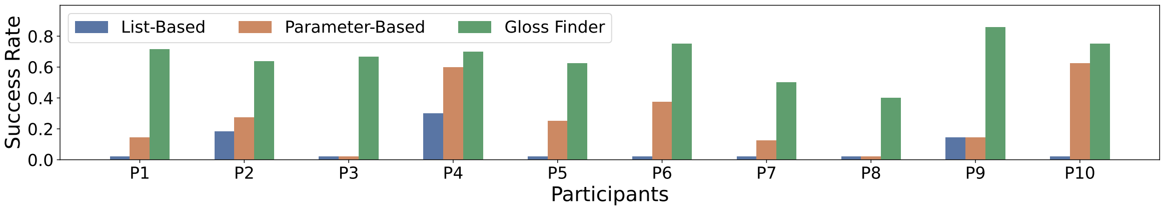


Figure 3: The success rate for each participant P1-P10 in reaching the correct sign during the trials.

the purpose of search for glosses. They are mainly of two categories:

- **List-based Dictionary:** All supported glosses were listed in their alphabetic order or grouped by category. We adopted widely used Signing Savvy (Signing Savvy, 2021) as the default one for participants, with the option of HandSpeak (Lapiak, 1995).
- **Parameter-based Dictionary:** The user was required to tick several parameters such as hand-shape to filter in the glosses. HandSpeak reverse dictionary (Lapiak, 1995) was adopted in our study.

3.2 Research Ethics and Recruitment

Ethical approval (Protocol 2021/375) was obtained from the Human Research Ethics Committee of The Australian National University. Each participant provided written informed consent.

Acknowledging that the primary users of this study was sign language leaners, the selection criteria were set to adults of any level of sign language experience. Particularly, experienced sign language users were favored if they were not using ASL. As such, we were able to collect direct feedback from experienced sign language users without forcefully asking them to pretend knowing the target signs. All participation was voluntary to encourage both positive and negative feedback.

3.3 Interview Process

The evaluation of the system is conducted in a form of interview with the participants to collect both quantitative and qualitative results. Each participant is guided to try each of the 3 target systems to search for specific signs. They started by experimenting with up to 6 signs from a determined set we constructed based on the consideration of the gesture diversity. The participant also chose a few signs from the vocabulary at will. Within the trial, they are encouraged to play around the systems for a few rounds to simulate the use of a dictionary. After some trials, they are asked to rate their satisfaction with the overall experience of interacting with the system, by taking into account both the quality of retrieved results and the support of the system to refine the search.

3.4 Participants

In total, 10 people participated in the study: 3 females and 7 males, and all in the age range of 20–40 years. The participants varied in their level of sign language experience. There were 3 intermediate sign language users with over 10 years of experience, of which 2 were attending professional jobs related to sign language interpretation. Additionally, there was 1 person having going through less than 1 year of systematic study, and 6 junior learners (a.k.a. beginners). All participants self-identified themselves as hearing, and were learning sign language for work, family members, or of their personal interest.

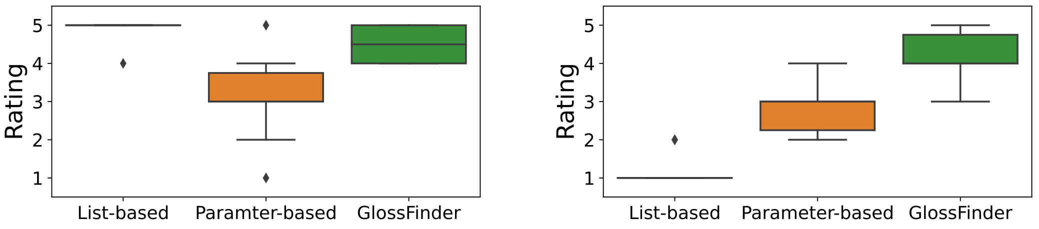
3.5 Evaluation Results

In this section, we summarize the ratings of the target systems from the three aspects as described below:

Ratings of learning to use the system The list-based dictionary and GlossFinder received higher ratings for easy to learn. A post-hoc analysis was conducted by Wilcoxon Signed-Rank Test to examine the significance of difference in pairwise comparison. Particularly, the corrected p-value were 0.0065 for LB-PB, 0.0103 for GF-PB, and 0.1025 for LB-GF [1]. We considered the difference is significant for the parameter-based method to the other two.

Noticably, the ratings for both the parameter-based system and GlossFinder rose after the trials. For the parameter-based system, the participants

[1] When it is not ambiguous, we will use the abbreviation of List-based (LB), Parameter-based (PB), and GlossFinder (GF) in reporting numeric results.



(a) Rating for learning the system

(b) Rating for using the system

Figure 4: Ratings our participants gave to the benchmark systems during the trials. The rating for learning the system was collected both before and after the trial to cope with the bias from the system introduction and long-term effects of the user getting familiar with the system.

had a tendency to be less frustrated when realizing they did not have to always specify all the parameters at once. For example, participant (P5) struggled on several signs about their location parameter especially for signs involving large or circular movements such as “family”. He decided to ignore the location in the following trials. In the meantime, the participants also showed more confidence in using GlossFidner as they got comfortable with using their camera. They also attributed the positive change to the capability of system to accommodate imprecise gestures and hand placement.

Ratings of using the system It was not surprising to see most participants quickly gave up using the List-based dictionary. Although some disagreement arose here, participants are giving a higher rating to GF in comparision to LB and PB. Similarly, Wilcoxon Signed-Rank Test was conducted to compare the pairwise significance of difference. The corrected p-values were 0.0069 for LB-PB, 0.0276 for GF-PB, and 0.0019 for GF-LB.

The Parameter-based system received controversial feedback among different groups. One of the intermediate signer had particularly used the system before and was reluctant to test it based on his previous experience. Junior learners showed a more optimistic view towards exploring the parameters. However, they reported strong depression over the minimal feedback the system was giving to them, which often confused them about if they were getting close to the target.

The participants expressed their favor to GlossFinder justified by the quality of results with the minimal effort. As pointed out by the participant P6, when he was searching for a sign (e.g., “travel”, performed by circulating a hand with two fingers bent forwards) in practice, he might not have been certain if the two fingers should have been pointing towards the front or up. This type of a local change led to less confidence in selecting the parameters. GlossFinder was less demanding on such preciseness. Some other testimonies focused on the interaction. As stressed for the parameter-based system, the users were not receiving feedback regarding their input. In contrast, they were able to compare against the gloss examples present in GlossFinder for possible matches. The participants said that the examples provided more than simply evidence for the correct match, but also guidance on how to proceed next if the target sign was not seen.

Success rate of search For quantitatively analyzing the system performance, we kept a record of the exact success rate for each trial of search during the interview, as the target sign was known to us. Detailed results can be found in Figure 3. For the list-based system, a success was defined as whenever the user clicks in the correct sign, no matter if they realized it is correct. It was extended for the parameter-based system to when the correct sign was in the first page of returned results (noticeably it was slightly favoring the system as the user might not be patient enough to examine all candidates even though they are certain with the correctness of parameters). For GlossFinder, we considered a success if the correct sign appears in the top-k results with $k = 12$ for that was the maximum number of videos to display in a common monitor resolution without scrolling. As shown in Figure 3, the success rate correlated positively with the user rating on usability and was clearly favoring the GlossFinder system. The average success rate for LB, PB and GF are 6%, 25% and 66% respectively. Most participants succeeded in locating the correct sign with 1 or 2 rounds of trials possibly by capturing a new video. Only the experienced signers were able to use the list-based system by relating the ASL sign here to the sign language they mastered and thus making a reasonable guess. In comparison, the results from the parameter-based system were relatively diverse, which unexpectedly was regardless of sign language experience. The testimony from participants of higher success rate suggested the method of only combining 2 parameters they were certain and brute force searching the candidates.

4 Conclusion

We construct, to the best of our knowledge, the first automatic sign dictionary digesting direct video capture as its inputs. Our user study validates the improved usability from the new system. The participants describe it as less demanding to learn in comparison to the existing parameter-based systems. Retrieved results are said to be more accurate and able to accommodate the varying video capture quality. Enriched results include example videos and explanations are agreed to largely help the user in correctly locating and refining the search. Overall, the reported success rate in reaching the searched sign is on average 66% from GlossFinder, significantly surpassing the benchmarks. We also conduct analysis to compare different views for presenting the results. It is favored by the participants for the system to include more examples of varieties even at the cost that less glosses can be shown in a single page. Our study strengthens the belief that the sign language dictionary design should be visual-based to imitate the practical form of actual sign language teaching and learning. We hope it can also motivate the related research to make sign language learning increasingly accessible to a broader community.

As one of the early attempts in building such system, we notice some limitations in the current study:

- The benchmark systems are comparatively weak, for which it is to blame the fact that sign language learners community is receiving insufficient support and no such stronger peers are public available. Existing systems are in majority made with voluntary contribution and limited in resource. While the incorporated benchmark systems are still receiving some positive feedback, stronger benchmarks are subject to encourage the participants to discover more places to improve in the current designs.
- The target audience of this study is set to general sign language learners, which is in concept a larger community covering DHH. We recruit people of both intermediate and junior level of sign knowledge to collect plausible data. Yet future research may be framed to be more customized for the DHH community. Space may still remain to improve based on their need.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*.

Oliver Alonzo, Abraham Glasser, and Matt Huenerfauth. 2019. Effect of automatic sign recognition performance on the usability of video-based search interfaces for sign language dictionaries. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 56–67.

LLC ASLSearch. 2009. Aslsearch.com.

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. 2021. Aligning subtitles in sign language videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11552–11561.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. *arXiv preprint arXiv:2203.04287*.

Jean Claude Corbeil and Ariane Archambault. 2006. *Merriam-Webster's Visual Dictionary*. Merriam Webster.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2735–2744.

Ralph Elliott, Helen Cooper, Eng-Jon Ong, John Glauert, Richard Bowden, and François Lefebvre-Albaret. 2011. Search-by-example in multilingual sign language databases. In *2nd Intl. Workshop on Sign Language Translation and Avatar Technology (SLTAT)*.

David Goldberg, Dennis Looney, and Natalia Lusin. 2015. Enrollments in languages other than english in united states institutions of higher education, fall 2013. In *Modern Language Association*. ERIC.

Jiahui Hou, Xiang-Yang Li, Peide Zhu, Zefan Wang, Yu Wang, Jianwei Qian, and Panlong Yang. 2019. Signspeaker: A real-time, high-precision smartwatch-based sign language translator. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–15.

Jolanta Lapiak. 1995. Sign language • asl dictionary | handspeak. https://www.handspeak.com/.

Vera Laska. 1993. The macmillan visual dictionary (book review). *International Journal on World Peace*, 10(4):107.

Seungyon Lee, Valerie Henderson, Harley Hamilton, Thad Starner, Helene Brashear, and Steven Hamilton. 2005. A gesture-based american sign language game for deaf children. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pages 1589–1592.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Dongxu Li, Chenchen Xu, Liu Liu, Yiran Zhong, Rong Wang, Lars Petersson, and Hongdong Li. 2021. Transcribing natural languages for the deaf via neural editing programs. *arXiv preprint arXiv:2112.09600*.

Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. 2020b. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045.

Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. 2020c. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6205–6214.

Liliane Momeni, Gul Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2020. Watch, read and lookup: learning to spot signs from multiple supervisors. In *Proceedings of the Asian Conference on Computer Vision*.

Carol Neidle, Ashwin Thangali, and Stan Sclaroff. 2012. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*. Citeseer.

Brian Scassellati, Jake Brawer, Katherine Tsui, Setareh Nasihati Gilani, Melissa Malzkuhn, Barbara Manini, Adam Stone, Geo Kartheiser, Arcangelo Merla, Ari Shapiro, et al. 2018. Teaching language to deaf infants with a robot and a virtual human. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Jacob Schioppo, Zachary Meyer, Diego Fabiano, and Shaun Canavan. 2019. Sign language recognition: Learning american sign language in a virtual environment. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

LLC Signing Savvy. 2021. Signing savvy | asl sign language video dictionary.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355.

Richard A Tennant, Marianne Gluszak, and Marianne Gluszak Brown. 1998. *The American sign language handshape dictionary*. Gallaudet University Press.

Haijing Wang, Alexandra Stefan, Sajjad Moradi, Vassilis Athitsos, Carol Neidle, and Farhad Kamangar. 2010. A system for large vocabulary sign search. In *European Conference on Computer Vision*, pages 342–353. Springer.

WHO. 2021. Deafness and hearing loss.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

A Testimony and Discussion

We extend the semi-structured open discussion with the participants on the general thoughts about sign language dictionary.

What should be the input format for a sign language dictionary?
Strong preference is given by the participants towards a rather simple input format to use the dictionary. One surprising finding is that despite experienced users are able to understand better the specification required in the parameter-based system (e.g. to choose the correct handshape and hand location), they also show stronger concerns over the capability as less-structured and more-composite signs are taken into scope (e.g. signs involving different movements from both hands). Junior sign learners show neutral altitude to learning the parameter-based system but it is also stressed by them the challenge they faced in understanding the parameters without further instruction.

> It is hard to know the clear definition of these parameters. Like now I feel I have to try each of these as they all look legitimate. (P5)

> Imaging a sign I don't know, maybe I won't really remember the gestures exactly but just a rough idea. (P6)

The participants express their favor towards the video capture used by GlossFinder for its nature correspondence to how people learn from peers. One improvement suggested is on the fact that both the interface and input require hand movements, i.e., they have to click the button and place the hand back to perform the sign. The future design may incorporate other UI elements to help the user focus on performing the sign. Examples include gesture-based UI input and foot controller.

> It would be cool if I can get all things done just by gestures. (P10)

One thing raised by the experienced signer P2 is:

> What are the things the system is looking at? ... Is it reading my lip as well? (P2)

As discussed in Section 2, the lip input is purposely dropped because of the inconsistency of quality and we want to prevent the model from accidentally learning to overfit to the lip-reading instead of the gesture. However, P2 pointed out that lip movement can be crucial in determining some of the signs, potentially a factor to consider in future development.

What should the dictionary show as the result?
Consensus is made by the participants on the advantage of displaying example videos:

> I can guess the meaning of some of these signs as I know them in another (sign language). I have no idea what other people would do if they only see the glosses in English (text). (P2)

> It becomes immediate now I know I find it. (P9)

In the meantime, future work is suggested on improving the order among variety examples for each individual gloss in the example-centric view. The matched glosses are ranked by confidence but the examples within each gloss are not. The result can be more reasonable if it can ensure the matched varieties to appear higher.

In addition, GlossFinder retrieves a fixed number of examples each time. It is argued that the number should adapt to cases for a clearer view.

> Some signs have many similar examples and it is good you show all of them. Just I feel like there may not always be so many similar glosses to show, and you see the later examples in the result become less meaningful. (P10)

As a language learning tool, what else should a dictionary have?
Since the primary audience of dictionary is language learners. We encourage the participants to think what can be improved from this perspective. A major point raised is that the dictionary may provide guidance on improving their sign. Even in case the dictionary retrieved the correct gloss, it is said:

> If you can put a confidence score for my recording, it sort of tells if I now remember it correctly. (P4)

A more sophisticated design may incorporate more instructions than the a score.

Maybe the dictionary can indicate the problems as I perform it. I see some difference in my form compared to that professional. (P4)