Gamification of RF Data Acquisition for Classification of Natural Human Gestures

Emre Kurtoğlu[1], Kenneth DeHaan[2], Caroline Kobek Pezzarossi[3], Darrin J. Griffin[4], Chris Crawford[5], Sevgi Z. Gurbuz[1]
[1]*Dept. of Electrical and Computer Engineering, The University of Alabama*
[2]*Dept. of American Sign Language, Gallaudet University*
[3]*Dept. of Psychology, Gallaudet University*
[4]*Dept. of Communication Studies, The University of Alabama*
[5]*Dept. of Computer Science, The University of Alabama*

2024 IEEE Radar Conference (RadarConf24) | 979-8-3503-2920-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/RADARCONF2458775.2024.10548148

***Abstract*—In recent years, there have been significant developments in radio frequency (RF) sensor technology used in human-computer interaction (HCI) applications, specifically in areas like gesture recognition and more broadly, human activity recognition. Although extensive research has been conducted on these subjects, most experiments involve controlled settings where participants are instructed on how to perform specific movements. However, when such experiments are conducted on sign language recognition they lack capturing dialectal and background-related diversities. In this work, we explore the differences in RF datasets acquired under controlled experimental settings and in free form environments where users were not constrained by the experimental instructions and limitations. We show that directed (i.e., controlled) data acquisition approaches result in over-optimistic performances which do not perform well on naturally acquired data samples in a real-world use case. We evaluate different approaches on generating synthetic samples from directed dataset, but show that such methods do not offer much benefit over collecting natural data. Therefore, we propose an interactive data acquisition paradigm through gamification. We show that the proposed approach enables the recognition of American Sign Language (ASL) in real-world settings by achieving 69% accuracy on 29 words.**

***Index Terms*—Micro-Doppler spectrogram, American Sign Language, RF sensors, deep neural networks, multi-modal**

I. Introduction

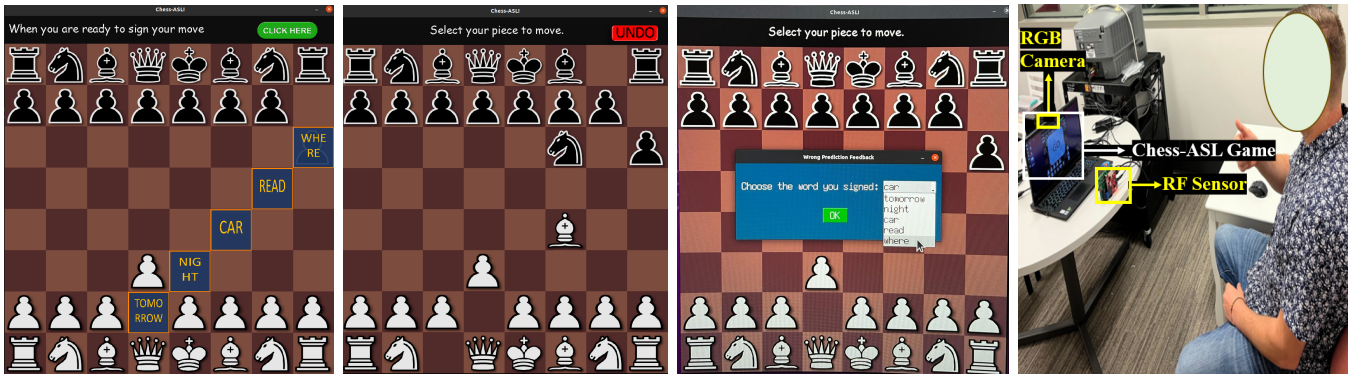
In recent years, there have been significant advancements in human-centered interactive technologies. While these developments make everyday life easier and more connected, Deaf community has not been able to take advantage of the speech recognition-based technologies. Ongoing studies on Deaf-centric research employ various sensors including video cameras [1], wearables [2], depth-augmented cameras (RGB-D) [3] and radar [4], [5].

Radio-frequency (RF) sensor present distinct advantages for sign language recognition (SLR) where other modalities have drawbacks. They are non-contact devices which can do ambient monitoring 24/7 without needing to be explicitly turned on/off. Moreover, they can operate seamlessly in dark and without being affected by people's clothing or skin colors. Also, they can measure physical variables in the scene such as velocity, distance and angle without needing to rely on certain estimation techniques used in video-based sensing applications. These features make radars suitable sensors for indoor applications.

However, as opposed to video-based applications, the amount of publicly available RF datasets are very limited and since there exists a wide range of RF sensors with different waveforms and operation characteristics, most of the time, the available datasets are not compatible with the sensor researchers are working with. When it comes to SLR tasks, more challenges are introduced. Cultural and background related differences in signing, regional dialects and fluency in signing are just to name a few. These diversities can result in significant change in data distribution. A deployed, real-world SLR system should be able to learn and adapt to these changes.

Conventional way of collecting data in a lab environment with strict experimental limitations and assumptions result in biased data which do not take real-world conditions into account and yield over-optimistic results. Similarly, collecting sign language data in a lab environment from participants whose primary language is not American sign language (ASL) has many sociological and technical drawbacks. For instance, not being aware of cultural nuances, having biased data for certain region, signing only a particular version of a sign and unnatural posture during signing can be listed as exemplary issues. Considering wide range of variety and complexity in sign language expressions, existing ASL datasets (in any sensing modality) are also not comprehensive enough to represent signing of different cultures, ethnicities and regional accents.

In this work, we both qualitatively and quantitatively show the differences of directed data (acquired under controlled experimental settings) and natural data (without experimental limitations or constraints). Several methods to overcome this challenge are evaluated and benchmarked. Finally, we introduce an interactive way of acquiring and recognizing sign language via gamification. We developed a chess game controlled with ASL where users can interact with the interface without needing to an external operator. The proposed system acquires multi-modal (video + RF) data, processes it, annotates the data, runs the chess engine and makes prediction on user data to move the pieces on the board.



(a) Highlighted possible moves with assigned words for a selected piece (left Bishop).

(b) UNDO button for correcting wrong predictions.

(c) Feedback GUI to choose the correct sign.

(d) ChessSIGN game environment setup.

Fig. 1: Gameplay pictures of the ASL-controlled chess game.

II. Gesture-Controlled Chess Game

Chess is a widely enjoyed strategic game which has drawn attention from individuals across various age ranges and backgrounds. It is compatible with our proposed interactive data acquisition method since it inherently has moderate pace, providing users ample time to contemplate and choose their moves. Therefore, it mitigates the limitations associated with real-time processing, providing sufficient time for local storage, data transfer, signal processing, and model inference. Additionally, the adaptable nature of chess enables the incorporation of extra features to capture intricate signing sequences and gather user feedback through a compact pop-up window. This capability empowers users to annotate their data during the gameplay, reducing the need for extensive subsequent quality control efforts.

When transitioning the data collection process into a gaming environment, various considerations arise which are not observed in controlled experiments. These include the need to design the game to be enjoyable and to ensure that any additional workload for self-annotation is not overly burdensome or disruptive to the extent that users become disinterested or frustrated with the interface.

Furthermore, it is crucial to reduce computational overhead associated with data processing as much as possible to prevent undesired delays in the game, which could negatively impact the user's gaming experience. Additionally, the predictions generated by the game control model should be sufficiently accurate to minimize instances where users need to undo their moves or feel like they are making mistakes or lack the necessary skills to play the game effectively.

The developed interactive ASL-enabled chess game is designed to collect data from both an FMCW radar and an RGB camera simultaneously. In our preliminary pilot iteration, the game control relies on predictions derived solely from video data. To mitigate possible user frustration resulting from misclassifications, we utilized a publicly accessible video-based ASL dataset to train the initial game control model, a process elaborated on in the subsequent section.

TABLE I: ASL signs utilized in the ASL-enabled chess game.

WATER	YES	BOOK	SLEEP	CAR	HELLO
HOME	READ	TIME	BETTER	DRINK	TOMORROW
SEE	HOT	BED	WHY	WHERE	LIKE
PLEASE	HAVE	MORNING	FINE	GO	NIGHT
CAN	TABLE	THERE	FINISH	HATE	

A. Video ASL Dataset for Pilot Deployment

To develop our initial video-based control model, we utilized Google's Isolated Sign Language Recognition (GISLR) dataset [6]. This dataset encompasses 250 fundamental concepts/vocabulary-based signs, representing the initial signs taught to infants in any language. The dataset comprises approximately 100,000 videos, with around 400 samples per class, featuring isolated sign expressions performed by 21 Deaf participants proficient in ASL. The corpus is constructed using hand, pose and facial landmarks generated by the MediaPipe Holistic pipeline, incorporating distinct models for each component. In this study, a subset of 29 signs is employed, corresponding to the maximum number of different positions the most mobile chess piece, the Queen, can move. These signs are selected from the GISLR dataset and are used to control the movement of the game pieces. A list of signs utilized is given in Table I.

B. Video Model for Pilot Study

The GISLR dataset was first provided in a hackathon organized by Google on Kaggle. Hoyeol Sohn [7] claimed the first place with a DNN model consisting of 1D-CNN and Transformer sub-networks. The model uses padding and truncation to handle varying-length input sequences. This approach satisfied the required inference speed while enabling usage of relatively larger models. For this study, we utilized the proposed model by modifying the output layer to have 29 neurons for the words of interest.

C. ChessSIGN Game Design and Play

The graphical user interface (GUI) of the developed ChessSIGN game comprised of two regions: chess board with pieces on it and the top banner where instruction prompts and interaction buttons are displayed. The game starts similar to a conventional chess game by clicking on a piece to move. Upon piece selection, all the possible moves for the selected piece are highlighted and each highlighted square has a randomly selected word among 29 signs written on it. Figure 1a shows the textual prompts when a piece is selected. Users can click on different pieces to explore where each piece can move. This feature becomes especially useful for inexperienced users. Once user decides where they want to move the piece, they click on the green "CLICK HERE" button located on the top right corner of the screen. This button starts a countdown from 3 to 0, and prompt a "GO" command when the countdown ended and starts the data recording for both camera and RF sensor for 3 seconds. User articulates the sign during this time frame and when the recording is ended, the acquired data sample is passed into the prediction model. The prediction model outputs the word with the highest probability, and the piece gets moved to the position where that word is located. After the move, a red "UNDO" button is displayed on the top right corner of the screen as shown in Figure 1b. If the prediction is correct, user selects another piece and continues to play. Otherwise, they click on the "UNDO" button which cancels the last move and pops-up a feedback window that gives user option to choose the correct word they actually signed from a drop-down menu. This feature enables users to self-annotate their own data, eliminating the labor-intensive labeling procedure after the data collection. The game history is logged into a text file along with the mispredicted file names to allow further offline analysis of the acquired data. The recorded samples are also uploaded to a cloud platform via its own application programming interface (API) in the back-end for data storage and remote access purposes.

The ChessSIGN game inherently preserves all the functionalities and rules of a regular chess game. It's main difference lies in the way game is controlled and other added functionalities like operating sensors, collecting data and running the prediction model. These automated capabilities eliminates the need for an operator to be present at the gaming area and an annotator to manually watch and label the ground truth classes.

D. Acquired ASL Datasets

1) Directed RF ASL Dataset: This dataset is acquired in 2022, in a laboratory environment under controlled settings where participants were instructed to sign a particular version of the sign by prompting them the sign on a monitor. Participants were seated on a chair facing towards co-located radar and the monitor. The RF sensor was placed approximately 0.9m above the ground and 1.5m away from the participants. Upon watching the exemplary video, they signed the same articulation of the word. Both hands were placed on a resting position on the knees before the recording started and retracted back to the original position when the signing is finished.

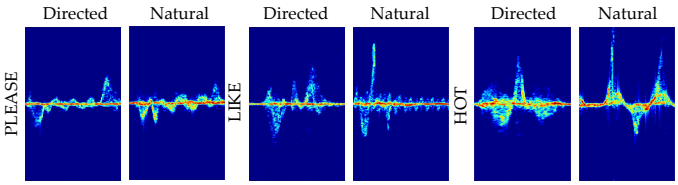


Fig. 2: μD spectrograms of directed and natural ASL samples for the signs PLEASE (left), LIKE (center) and HOT (right).

This way, it is ensured that all the participants articulated the same version of the sign and the cultural or background-related variances are minimized. 19 Deaf/HoH participants from Gallaudet University (the only university in the U.S. whose primary language of instruction is ASL and tailored for Deaf/HoH students) and 4 participants from The University of Alabama attended the study. While 21 of the participants were Deaf, two of them are Child-of-Deaf-Adults (CODAs) who were fluent ASL signers. Studies in both locations were conducted under the same experimental settings with the same operators. 110 signs from ASL-Lex database [8], including verbs, nouns and adjectives were selected based on their kinematic variability and usage frequency. Around 40 samples per class are acquired which adds up to 4,455 samples in total.

2) Multi-Modal Interactive ASL Dataset: This dataset is acquired during the gameplay of ChessSIGN game in 2023. The participants were seated in front of the RF sensor and the laptop in which game was running as shown in Figure 1d. An RGB camera was also utilized to capture the signing videos. Both sensors were triggered simultaneously for synchronized data capture via the game. In total, 23 Deaf participants have attended the study at Gallaudet University. Note that the participants attended to this study are different than those participated in the directed ASL study. Around 37 samples per class were collected for 29 signs which added up to 1,078 samples, in total.

E. RF Sensor and Data Pre-Processing

In this work, Texas Instrument's AWR2243BOOST radar evaluation module coupled with DCA1000EVM data capture card are used for raw data acquisition. The RF sensor is a frequency-modulated-continuous-wave (FMCW) multiple-input-multiple-output (MIMO) radar operating at 77 GHz with a maximum of 4 GHz bandwidth.

Moving targets in radar field-of-view (FoV) cause frequency shift in the received signal. The micro motions generated by arm strokes and finger movements result in micro-Doppler (μD) [9] shifts in the received signal. Time-frequency analysis of the received signal reveals the unique μD pattern of each sign and motion. This transformation is called μD spectrogram and can be computed as the square modulus of the Short-Time Fourier Transform (STFT) of the input signal. In this work, μD spectrograms are generated with sliding windows of length 256 pulses with the overlapping region of 200 pulses.

TABLE II: Statistical comparison of velocities of directed and natural signing.

Data Type	Avg. Max. Velocity	Var. of Max. Velocity	Avg. Min. Velocity	Var. of Min. Velocity
Directed	2.58 m/s	1.06 $(m/s)^2$	-2.46 m/s	0.89 $(m/s)^2$
Natural	2.61 m/s	1.59 $(m/s)^2$	-2.28 m/s	1.03 $(m/s)^2$

III. Directed versus Natural Data

In this section, we compare the RF ASL datasets acquired under controlled experimental settings (i.e., directed) which is conventional way of collecting data and during the game play with natural interactions. We first address the issue of differences between two approaches and qualitatively show how directed experiments fail to represent the nuances of natural signing. Next, we show how lack of capturing these features deteriorate the training procedure and cause over-optimistic results for real world applications.

A. μD Spectrogram Comparison

Figure 2 shows μD signatures of directed and natural samples for the signs PLEASE, LIKE and HOT. As can be seen from the figure, there exist significant differences for the same signs, and these differences are not slight changes in time span or Doppler bandwidth, but rather major changes in number of arm strokes, negative and positive Doppler peaks. By watching the corresponding camera videos for μD spectrograms, differences in the way participants sign can be observed. For instance, for the word PLEASE, the participant in natural signing moves her arm towards her chest in two steps with a short pause in-between instead of one movement. This creates two consecutive negative peaks at the beginning of the sign instead of one. Also, the peaks at the beginning and ending of the signing caused by the major arm movements towards and away from the chest have lower peaks than the directed case. For the word LIKE, the participant in natural settings study shakes her hand after the signing is finished which causes jittering effect at the lower Doppler frequencies. Finally, for the word HOT, the participant in natural settings repeat the sign twice resulting in two positive Doppler peaks. This case was seen in several other signs as well where the participants sign the word multiple times perhaps with the purpose of "convincing" the system to the articulated sign. Considering most of the users are and will be unfamiliar with working principles of radar, these unexpected abnormalities or dialect-related diversities should be handled by the system to provide a satisfactory user experience.

B. Comparison of Velocities

Evaluation of velocity profiles of directed and natural ASL samples can give information regarding diversity of the datasets. Table II presents average maximum and mean velocities along with their variances for directed and natural datasets. From these results, it can be inferred that although the two datasets have almost identical average maximum velocity, the variance of natural samples are significantly higher than the directed ones (i.e., 1.59 $(m/s)^2$ vs. 1.06 $(m/s)^2$). Similarly, in minimum velocities, natural samples have a greater variance (1.03 $(m/s)^2$) than the directed samples (0.89 $(m/s)^2$).

TABLE III: Performance comparison of different training methods and datasets.
(Note that no natural signing data are used in the training phase of Exp. 5).

Exp. ID	Training Data	Testing Data	Modality	Model	Acc. %
1	GISLR	GISLR	Video	1D-CNN + Transformer	92.3
2	GISLR	Natural ASL	Video	1D-CNN + Transformer	48.2
3	Directed ASL	Directed ASL	RF	2D-CNN + MLP	68.9
4	PhGAN(Dir. ASL)	Directed ASL	RF	2D-CNN + MLP	100
5	PhGAN(Dir. ASL)	Natural ASL	RF	2D-CNN + MLP	9.6

C. Impact on Model Training

The difference in data distribution of directed and natural signing samples are more stressed when a prediction model is trained on directed data and tested on natural samples. It is found that the distribution difference of the datasets collected with two different approaches are so high that the model trained with directed data is not able to recognize natural signing at all. This phenomenon is observed for both camera and radar data.

First, it should be noted that the model performance can be evaluated in two different ways: "in-game" accuracy and offline accuracy on 29 signs. In-game accuracy refers to the accuracy experienced by the user during the course of game where only certain number of signs are presented to the user and the model predicts the most likely one. For instance, a Knight can move up to eight different positions at once. This limits the number of classes the model makes prediction on. After the game is concluded, an offline prediction can be made on 29 classes and the true performance can be obtained. Therefore, in-game accuracy results are typically higher than the offline prediction results.

1) Video-Based Model: The initial video-based model trained with GISLR dataset yielded 92.3% accuracy on the testing portion of the dataset. However, during actual chess game play, the in-game accuracy experienced by the users was only 76.62%. Moreover, after the study ended, when the acquired data is tested on 29 signs it was only 48.24% which is significantly lower than the testing accuracy attained on the GISLR dataset. This results demonstrates the limitations of available datasets acquired in controlled experimental settings when they are deployed for a real-world application. They do not well-represent the nuances and features of sign languages emerge in natural use cases.

It is observed that more kinetic signs have higher confusions like HOT, FINISH, FINE, HELLO, GO and HAVE. One possible reason for these signs to have more confusion could be that

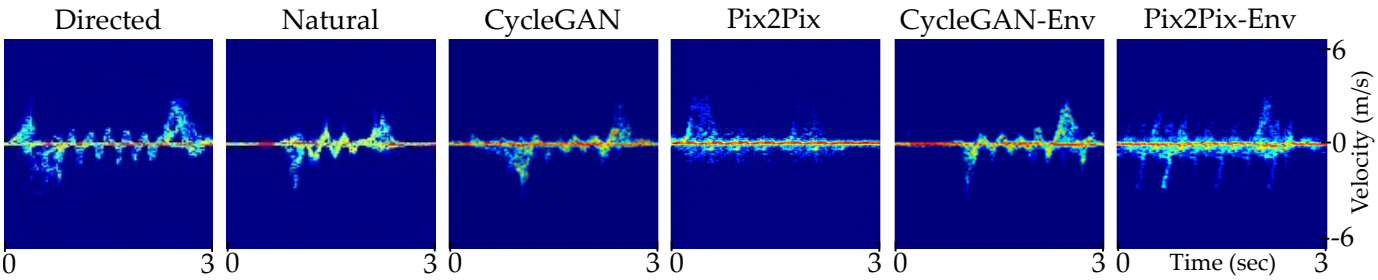


Fig. 3: μD signatures of the PhGAN-synthesized directed and natural samples, and benchmarking of transformed samples generated by CycleGAN, Pix2Pix, CycleGAN-Env and Pix2Pix-Env models.

while video is very effective in capturing spatial variance, it is not as effective in capturing depth and temporal variance. This drawback can be be compensated by RF sensors which are more capable of capturing dynamics of signing instead of shapes.

2) Radar-Based Model: A similar phenomenon is observed in the radar prediction model. Radar prediction model consists of 4 2D convolutional neural network (CNN) blocks followed by 2 multi-layer perceptrons (MLPs). The radar model trained with directed real data samples yielded 68.9% accuracy on directed testing samples. It should be noted that there is a significant difference in number of real samples of video and radar data per class (i.e., 40 vs 400). In a prior work, physics-aware generative adversarial network (PhGAN) [10] method is found to be very effective in augmenting the training data by generating kinematically accurate synthetic samples. Hence, in this work, PhGAN model is used to increase the number of directed samples, and 500 synthetic samples are generated. Inclusion of PhGAN-synthesized spectrograms in the training stage boosted the accuracy of the RF model on directed dataset to 100%.

Although this result looks great, this model completely fails to recognize natural ASL signing when tested on the data acquired during the chess game play, yielding only 9.56% accuracy for 29 signs. This result is much worse than the performance drop observed in the video which dropped from 92.3% to 48.2%. There can be several reasons why radar is more affected when compared the video. First, even though the number of training samples were increased for RF data through PhGAN method, the video model was trained with significantly higher number of real samples when compared to the radar model. Secondly, there are major differences in the way directed ASL and natural ASL signs are articulated which can potentially change the kinematic and temporal progression while spatial similarities are still preserved. Since radar is more sensitive to radial and temporal changes rather than target shapes or spatial features, its performance is more dramatically affected.

IV. Interactive Learning of Natural ASL

The results presented in the previous section show the challenges in recognizing natural ASL and highlights the need of training models with naturally collected data. The proposed interactive system enables this in a sustainable, enjoyable manner. However, the question of how to best train the RF model is yet to be answered. This section tackles this problem and compares different approaches.

A. Domain Transfer from Directed to Natural ASL

One way of mitigating the difference between directed and natural ASL data could be to utilize domain adaptation techniques. They can be used to bridge the gap between the two distributions. Instead of training the network with directed samples, they can be transformed to natural-like samples and could be used for training afterwards. In this study, in addition to using PhGAN for sythetic data generation from small amount of natural samples, we assess the efficacy of two domain adaptation methods from directed data: CycleGAN [11] and Pix2Pix (P2P) [12]. In this study, these networks are basically trained to map the given directed ASL sample to more natural-like ASL samples. Architectural details of these methods are omitted for brevity.

In addition to vanilla CycleGAN and Pix2Pix models, we developed a modified versions of these models (i.e., CycleGAN-Env and Pix2Pix-Env) where a physics-based loss term is included to generate kinematically more accurate samples. These versions extract the upper and lower envelopes of the μD spectrograms using the percentile technique [13]. The introduced physics-aware loss, $\mathcal{L}_{\text{ph}}$, is computed as the mean-square error (MSE) between the envelopes of the target and the generated μD spectrograms. Then, the total loss of the GANs, $\mathcal{L}_{\text{GAN}}$, can be written as

$$\mathcal{L}_{\text{GAN}}(G, D_N, D, N) = \mathbb{E}_n[\log(D_N(n))] + \mathbb{E}_d[\log(1 - D_N(G(d)))] + \lambda \mathcal{L}_{\text{ph}}, \quad (1)$$

where G is the generator, D_N is the discriminator for natural domain, D and N are directed and natural domain samples and λ is the weighting factor for $\mathcal{L}_{\text{ph}}$. The first and the second terms represent the discriminator and the generator losses, respectively.

Figure 3 shows a sample spectrogram for visual comparison of PhGAN-synthesize directed and natural μD spectrograms, and the transformed samples generated by CycleGAN, Pix2Pix, CycleGAN-Env and Pix2Pix-Env methods. It can be qualitatively observed that vanilla CycleGAN does a better

TABLE IV: Final classification results of VGG-16 for RF data of natural ASL.

Method	PhGAN	CycleGAN	CycleGAN with Env.	P2P	P2P-Env
Acc. (%)	**69.14**	62.04	62.35	61.73	60.49

job in resembling directed sample to the natural one when compared to Pix2Pix. Although CycleGAN produced the first and the last peaks very well, middle peaks are not well-represented. CycleGAN-Env, on the other hand, seems to be able to successfully create those peaks with a strong signal strength. This shows the efficacy of the introduced physics-aware loss term in the data model training stage. A similar phenomenon can be observed when Pix2Pix output is compared with Pix2Pix-Env. While vanilla Pix2Pix is not able to reproduce most of the peaks, Pix2Pix-Env is able to replicate periodic peaks with high signal strength.

B. Fine Tuning with Synthetic Natural ASL

The small amount of real natural data can be used to generate a large amount of synthetic samples. As more samples are produced using domain adaptation or synthetic data generation techniques, deeper models can be trained. In this study, we utilize 16-layer CNN architecture of VGG-16 [14] pre-trained with ImageNet database [15] weights. While such pre-trained networks are unaware of high level RF data features, they are well-trained on common primitive features such as edges, lines and corners. Fine tuning them with the RF data enables them to learn high level RF data features as well - making them more powerful when compared to shallow networks. In order to make a fair comparison amongst different methods, we chose VGG-16 as the benchmarking network and trained different versions of it using each method. 30% of the natural signing data was always spared for testing and the remaining 70% was used to drive different methods to generate synthetic samples. Table IV presents the accuracy results obtained using those methods. It can be observed that PhGAN-synthesized natural samples yield the best accuracy of 69.14% which is around 7-8% better when compared to domain adaptation techniques. This results show that data adaptation methods from directed ASL under-performed when compared to simply synthesizing data from natural ASL.

V. Conclusion

In this work, we propose an ASL-enabled interactive chess game, ChessSIGN, as a new way of acquiring natural multi-modal (video + RF) data. We show that the traditional way of collecting data under controlled experimental settings result in biased data which do not well-represent linguistic and dialectal properties of natural signing. Therefore prediction models trained with such datasets yield over-optimistic results and fail to recognize natural data when deployed in a real-world environment without experimental limitations/assumptions.

Efficacy of different domain adaptation techniques to transform directed data to more natural-like data is also evaluated and found to be under-performing when compared to acquisition of real natural data. Therefore, this work underscores the importance of acquiring natural data in an interactive manner. The proposed system achieves 69% accuracy for 29 signs even with a small amount of real data. Inclusion of other sensing modalities to the system is yet to be explored.

Acknowledgment

This work was funded in part by the American Association of University Women (AAUW) via a Research Publication Grant for Engineering, Medicine and Science, and by National Science Foundation Awards #1932547 and #2238653. Human studies research was conducted under UA Institutional Review Board Protocol #18-06-1271.

References

[1] D. Li, C. Rodriguez-Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1448–1458, 2019.

[2] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, "Wearable sensor-based sign language recognition: A comprehensive review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 82–97, 2021.

[3] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1750–1762, 2022.

[4] S. Gurbuz, A. Gurbuz, C. Crawford, and D. Griffin, "Radar-based methods and apparatus for communication and interpretation of sign languages," in *U.S. Patent 11,301,672 B2 (Invention Disclosure filed Feb. 2018; Provisional Patent App. filed Apr. 2019), granted April 12, 2022.*, October 2020.

[5] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. S. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, "American sign language recognition using rf sensing," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3763–3775, 2021.

[6] A. Chow, G. Cameron, M. Sherwood, P. Culliton, S. Sepah, S. Dane, and T. Starner, "Google - isolated sign language recognition," *Kaggle*, 2023.

[7] H. Sohn, "Google - isolated sign language recognition," https://www.kaggle.com/code/hoyso48/1st-place-solution-training, 2023, accessed: Sep. 13, 2023.

[8] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey, "Asl-lex: A lexical database of american sign language," *Behavior Research Methods*, vol. 49, no. 2, pp. 784–801, Apr 2017.

[9] V. Chen, *The Micro-Doppler Effect in Radar, Second Edition*. Artech, 2019.

[10] M. M. Rahman, S. Z. Gurbuz, and M. G. Amin, "Physics-aware generative adversarial networks for radar-based human activity recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 3, pp. 2994–3008, 2023.

[11] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, 2017.

[12] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016.

[13] P. V. Dorp and F. C. A. Groen, "Feature-based human motion parameter estimation with radar," *IET Radar, Sonar Navigation*, vol. 2, no. 2, pp. 135–145, 2008.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.