CSI-Based Sign Language Recognition Using CNN-GRU with Attention

Joy Saha
Department of Electrical and Computer Engineering
Email: jsaha2@albany.edu

I. Introduction

Sign language recognition is important for communication accessibility. Wi-Fi CSI provides a privacy-preserving, contactless modality capturing amplitude and phase variations across time, subcarriers, and antennas. These high-dimensional signals are challenging for traditional models. We propose a hybrid CNN-GRU with attention, data augmentation, and self-supervised pretraining to classify CSI gestures. Performance is compared against classical baselines (ANN, Decision Tree, Naïve Bayes, SVM, k-NN), and ablation studies quantify architectural contributions.

II. Dataset

We use three public datasets from [1]. Each sample is a $(200, 60, 3)$ tensor (200 time steps, 30 amplitude and 30 phase subcarriers per antenna, three antennas).

TABLE I
CSI sign language datasets

Dataset	# Labels	Repetitions	# Instances
Home	276	10	2760
Lab	276	20	5520
Lab150	150	10	7500

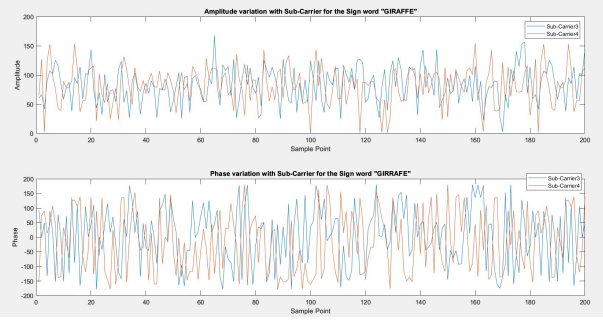

Fig. 1. Amplitude and phase variation across subcarriers for one antenna and subject.

III. Proposed Method

Our base CNN model [1] (Figure 2) uses 2D convolutions to extract spatial patterns from CSI across subcarriers and antennas. Dropout and batch normalization are applied to reduce overfitting.

The full proposed method extends this base model by adding a GRU layer to capture temporal dynamics across time steps and an attention mechanism to emphasize salient gesture-relevant sequences. The combined CNN-GRU-attention architecture allows the model to learn both spatial correlations and temporal dependencies in the CSI data. We also employ data augmentation (temporal warping and Gaussian jitter) and self-supervised pretraining to improve generalization.

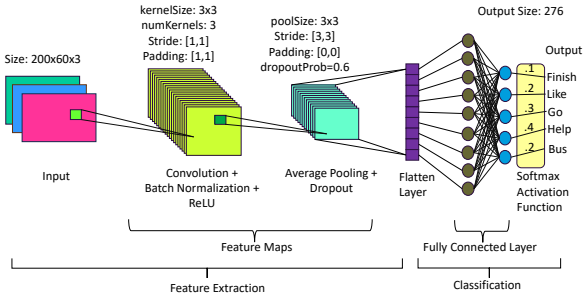

Fig. 2. Base CNN architecture.

IV. Baselines

We compare against: ANN, Decision Tree, Naïve Bayes, SVM, k-NN, using handcrafted statistical features (std, entropy, spectral energy, correlation).

V. Ablation

Components are systematically removed to assess contribution: CNN-only vs GRU-only vs CNN-GRU, attention removed, without augmentation, without pretraining, flattened vs 2D input. Accuracy, Training Time, and confusion matrices are reported.

VI. Explainability

Grad-CAM highlights which subcarriers contribute to predictions, improving interpretability.

VII. Conclusion

Hybrid CNN-GRU with attention can effectively classify CSI-based gestures. Results may generalize to broader human activity sensing.

References

[1] Y. Ma, G. Zhou, S. Wang, H. Zhao, W. Jung, "SignFi: Sign Language Recognition Using WiFi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–21, Mar. 2018.