Few-Shot Isolated Sign Language Recognition with Spatiotemporal SlowFast Prototypes

Michał Kalinowski and Bożena Kostek

Gdańsk University of Technology, Gdańsk, Faculty of Electronics, Telecommunications and Informatics, Multimedia Systems Department and Audio Acoustics Laboratory, Poland

Objectives

- Frame sign language recognition as a few-shot learning problem using a prototypical network.
- Extract spatiotemporal embeddings with a modified SlowFast CNN [1].
- Assess generalization to unseen classes on the LSA64 dataset.

Introduction

Sign language is the primary communication mode for more than 70 million deaf and hard-of-hearing individuals worldwide. However, sign languages are not globally standardized—over 300 distinct sign languages exist, often differing even between regions that share the same spoken language [2]. This linguistic diversity, combined with limited annotated video corpora, poses a significant challenge for automatic sign language recognition (SLR). A key difficulty is ensuring that models can generalize beyond the specific signs and signers seen during training.

Most existing SLR systems rely on large, fully supervised datasets and struggle to generalize to unseen signs or low-resource scenarios. Despite its importance, few-shot learning has been relatively underexplored in SLR, particularly for video-based recognition of isolated signs. To address this limitation, we explore a data-efficient approach that frames sign recognition as a few-shot learning problem. Our method integrates a modified SlowFast video encoder [1] under the prototypical learning paradigm to learn discriminative spatiotemporal representations and recognize sign classes from only a few examples, including classes not seen during training.

Highlights

- Recognizes unseen sign classes with very few examples.
- Strong feature clustering; errors mainly in visually similar signs.

GDAŃSK UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRONICS, TELECOMMUNICATIONS AND INFORMATICS

Main Sign Language Modalities

1. RGB: Captures rich visual details but is sensitive to lighting and background variation.
2. Pose/Skeleton: Compact joint representation but prone to occlusion errors and often misses facial grammar.
3. Optical Flow: Captures motion patterns and serves as helpful auxiliary input.
4. Depth: Helps disambiguate overlapping limbs and can improve detection accuracy in some cases.
5. Text/Glosses: Support translation and cross-modal alignment, helping reduce the modality gap.

Well-Established Architectures for SL Processing

1. CNN–RNN: Extract spatial features (CNN) and model temporal dynamics (RNN) for SLR and CSLR.
2. Transformers: Self-attention models for long-range context in CSLR/SLT.
3. GNNs: Skeleton graph modeling for SL processing, mostly isolated SLR.
4. Hybrid / Multistream: Parallel streams or combined modules (e.g., CNN + Transformer, RGB + skeleton) fuse complementary features for robustness and better generalization

The Proposed Pipeline

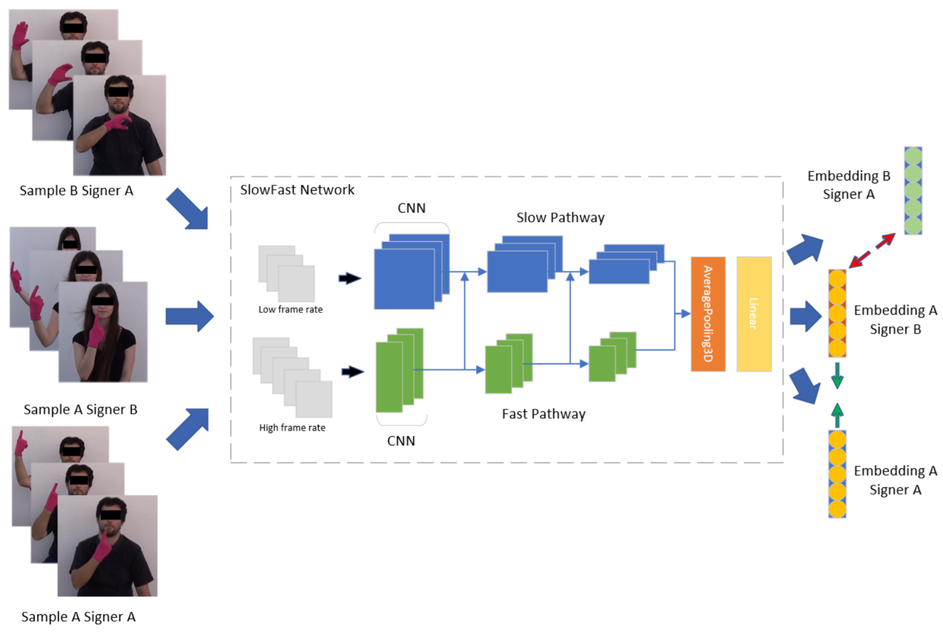

Figure 1:Illustration of the few-shot sign language recognition pipeline using a prototypical network with the SlowFast feature embedder [1]. The graphic shows the processing of video samples and the training process, where embeddings are learned to cluster same-class samples and separate different-class samples, enabling classification based on proximity to class prototypes.

The Proposed Method

1. Dataset & Processing: LSA64: 3,200 RGB videos, 64 signs, 10 subjects (5 reps), recorded with fluorescent-glove setup. Videos normalized [0,1], resized to 224×224, and padded for length. Split: 80/20 → 52 train / 12 val classes.
2. Training: We train for 30 epochs, with 100 episodes/epoch, using a custom episodic sampler. Due to memory limits, episodes use 4-way classification, with 3 support and 2 query samples per class. Training uses Euclidean distance and follows the prototypical network formulation, adapted for videos.
3. Experiments: Evaluation was performed on 1,000 episodes, testing 5-way and 10-way setups with support sizes 1, 5, and 10, and 15 queries per class, following few-shot evaluation protocol. For full test-set inference, embeddings were averaged to form class prototypes, and samples were classified based on nearest-prototype assignment. A confusion matrix summarizes per-class performance.

Results

Tested on 600 samples: 67 errors → 88.8% accuracy.

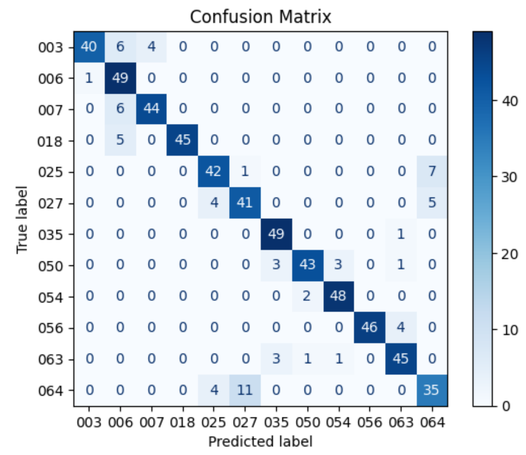

Figure 2:Confusion Matrix of Test Set Predictions. Visually similar signs (e.g., "025", "027", "064") show some misclassifications.

Conclusion

- Accuracy: 88.8% on few-shot isolated SLR.
- Feature extraction: Effective for unseen signs; most classes form distinct clusters.
- Challenges: Visually similar signs sometimes overlap, reducing separability.
- Limitations: Small subset of LSA64 tested; fluorescent gloves reduce real-world applicability.

Future directions:

- Evaluate on larger, glove-free datasets.
- Incorporate skeletal/structural cues (e.g., graph-based features).
- Explore margin-based loss and hyperparameter optimization to improve class separation.

References

[1] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020.

[2] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.

Acknowledgments

Research and Development (NCBR) project: "ADMED-VOICE Adaptive Intelligent Speech Processing System of Medical Personnel with the Structuring of Test Results and Support of Therapeutic Process," project no. INFOSTRATEG4/0003/2022".