FusionEnsemble-Net: An Attention-Based Ensemble of Spatiotemporal Networks for Multimodal Sign Language Recognition

Md. Milon Islam[1]*, Md Rezwanul Haque[1]*, S M Taslim Uddin Raju[1], Fakhri Karray[1,2]

[1]University of Waterloo, [2]Mohammad bin Zayed University of Artificial Intelligence

Introduction

- Sign Languages (SLs) are multimodal (manual & non-manual components).
- Crucial for deaf communities, especially in healthcare to bridge communication gaps.
- Existing SLR systems face challenges:
  - ✓ Difficulty capturing complex multimodal gestures.
  - ✓ Limited dataset diversity (signer demographics, environment, sensor modalities).
  - ✓ Privacy concerns with camera-based systems in healthcare.
  - ✓ Need for robust models that generalize across real-world scenarios.

Our Proposed Solution: FusionEnsemble-Net

Multimodal Data and Preprocessing

- **Input**: RGB video (visual info: handshapes, facial expressions, body posture) and Range-Doppler Map (RDM) radar data (motion info, privacy-preserving).
- Data synchronized, resized (224×224), and normalized.

Parallel Spatiotemporal Feature Extraction

- Utilizes an ensemble of four diverse spatiotemporal networks for robust feature learning.
- 3D ResNet-18
- MC3-18
- R(2+1)D-18
- Swin-B (transformer-based)
- Temporal modeling layers (LSTMs, transformer encoders, linear projections) capture dynamic sequences.

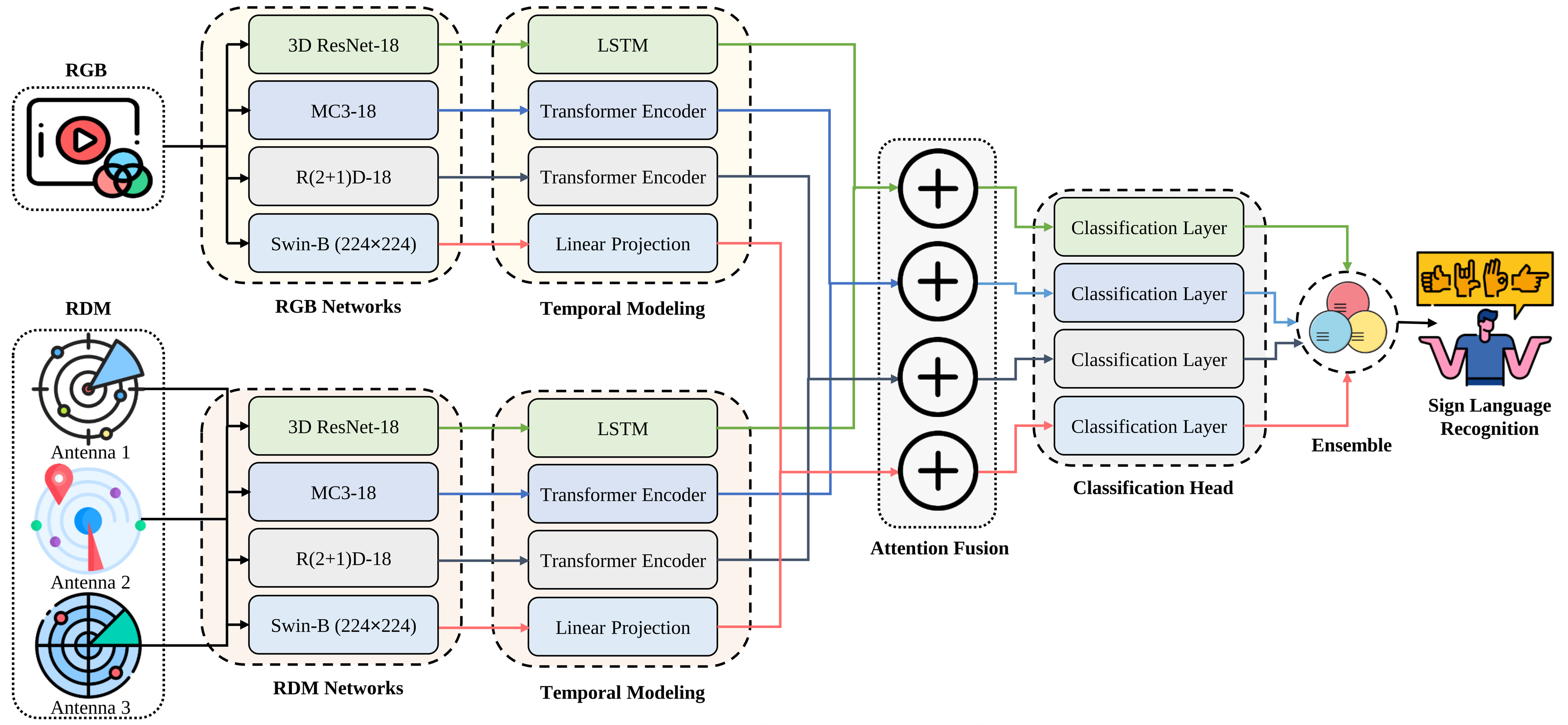
Attention-Based Feature Fusion

- Modality-specific temporal features are concatenated.
- A self-attention module dynamically re-weights visual and motion features to generate a single fused representation.

Ensemble Classification Head

- Each fused feature vector is passed to an independent classifier.
- Final prediction is an average of probabilities from all classifiers, enhancing robustness.

System Architecture



Dataset and Experiments

Dataset

- **MultiMeDaLIS:** A large-scale, multimodal dataset for isolated Italian Sign Language recognition in a medical context.
- **Content:** Contains 126 unique signs, including 100 medical terms and 26 alphabet letters.
- **Modalities Used:** We utilize the synchronized RGB video and RDM radar data.

Implementation Details:

- **Framework:** PyTorch.
- **Hardware:** Trained on two NVIDIA A6000 GPUs.
- **Optimization:** Used the AdamW optimizer with pre-trained weights from Kinetics-400 and ImageNet to leverage transfer learning.
- **Training:** The model was trained for 25 epochs, requiring approximately 44 hours.

Evaluation Metric

- Top-1 Accuracy on validation and test sets.

Results and Analysis

Methods	Modality	Valid	Test
SL-GCN [1]	RGB	-	97.98
SSTCN [1]		-	96.33
ResNet(2+1)D Optical Flow [1]		-	56.31
ResNet(2+1)D Frame [1]		-	97.29
ResNet(2+1)D Encoding HHA [1]	Depth	-	88.04
AutoTrans-RDMNet [11]	RDM	-	88.3
	3×RDM	-	91.7
	MTI	-	84.9
	3×MTI	-	86.1
	RDM+MTI	-	91.4
	3×RDM+3×MTI	-	93.6
3D ResNet	RGB+3×RDM	96.58	96.58
MC3		98.96	99.06
R(2+1)D		96.94	97.34
Swin-B		94.24	94.42
FusionEnsemble-Net		**99.37**	**99.44**

* HHA=Height, Horizontal disparity, Angle, and MTI=Moving Target Indications.

Conclusion and Future Work

Conclusion

- Our FusionEnsemble-Net sets a new SOTA accuracy of 99.44% on the MultiMeDaLIS dataset.
- Our diverse ensemble effectively fuses RGB and radar data for sign language recognition.

Future Directions:

- Extend to continuous, conversational sign language.
- Develop a lightweight version for real-time deployment.

Models and codes are publicly available

Link: https://github.com/rezwanh001/Multimodal-Isolated-ItalianSign-Language-Recognition.

[1] Caligiore et al.,“Multisource approaches to Italian Sign Language recognition,”CLiC-it 2024, Pisa, pp. 132–140.

[2] Mineo et al.,“Sign language recognition for patient-doctor communication,”IEEE RTSI 2024, pp. 202–207.