HySTA-NET : A HYBRID SPATIO-TEMPORAL ATTENTION NETWORK FOR RECOGNITION OF EMERGENCY SIGNS IN INDIAN SIGN LANGUAGE

Aisha Thameem (TKM21CS012), Bhagya A Jai (TKM21CS041), Fathima A (TKM21CS053), Uthara Sabu (TKM21CS138)

Project Guide : Dr. Shyna A, Department of Computer Science and Engineering,
TKM College of Engineering

Abstract

Effective communication is essential, enabling individuals who rely on sign language to express their thoughts and emotions seamlessly. Hybrid Spatio-Temporal Attention Network (HySTA-Net) is a novel deep learning framework for real-time recognition of emergency Indian Sign Language (ISL) gestures. By integrating CNN-based spatial feature extraction (using a modified VGG16), LSTM-based temporal modeling, and multi-head attention for enhanced feature representation, the system achieves high accuracy in classifying critical emergency signs. Tested on both public and custom video datasets, HySTA-Net demonstrates strong generalization, making it a promising tool for improving emergency communication for the deaf and hard-of-hearing community.

Objective

Develop a real-time ISL emergency gesture recognition system using deep learning to ensure accurate communication and enhance safety for the deaf and speech-impaired community.

Dataset

A publicly available dataset of 8 dynamic emergency ISL signs was used for training, while a custom dataset of 8 signs ,50 videos per sign from 7 signers was created for evaluation.

Accident

Call

Doctor

Help

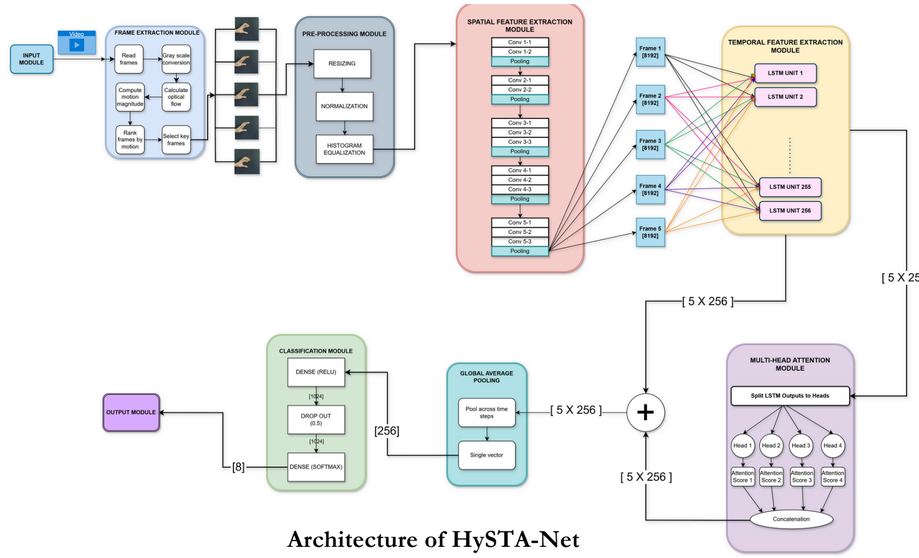
Hot

Lose

Pain

Thief

Methodology

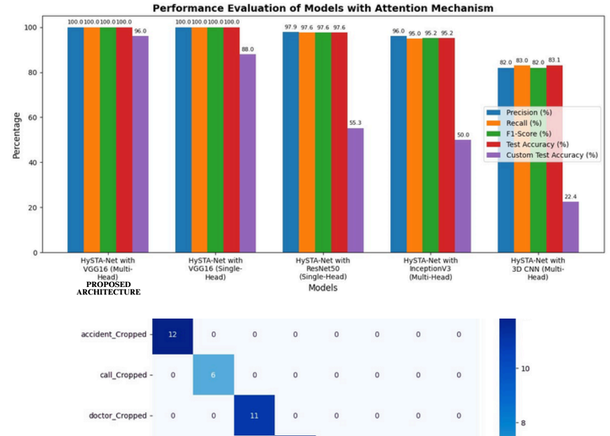


Architecture of HySTA-Net

- **Video Acquisition & Frame Extraction:** Capture raw videos and extract key frames using an optical flow-based method.
- **Pre-Processing:** Standardize frames via resizing (150×150), normalization, and histogram equalization.
- **Spatial Feature Extraction:** Use a modified VGG16 (wrapped in a Time Distributed layer) to derive spatial features from each frame.
- **Temporal Modeling:** Process sequential spatial features with a 256-unit LSTM to capture dynamic gesture evolution.
- **Attention & Residual Connection:** Apply multi-head attention (4 heads) and add a residual connection to refine temporal features.
- **Global Pooling & Classification:** Aggregate features via global average pooling and classify using dense layers with ReLU, dropout (0.5), and softmax activation.

Result and Analysis

Comparative evaluation shows that the proposed model HySTA-Net using VGG16 achieves 100% test accuracy and 96% custom test accuracy—outperforming state-of-the-art and other custom developed models. Graphs and a confusion matrix confirm that attention-enhanced models yield superior results, ensuring robust emergency ISL recognition. Confusion matrix analysis shows strong diagonal dominance, indicating minimal misclassifications across 8 emergency gestures



The results validate the technical approach and promises significant social impact by enabling timely communication for the deaf, hard-of-hearing, and speech-impaired communities in critical situations.

Conclusion

HySTA-Net effectively recognizes emergency ISL gestures by combining spatial and temporal deep learning techniques, addressing the crucial need for accessible communication in emergencies for the deaf and hard-of-hearing community.

References

[1] Q. M. Areeb, R. Alroobaea, M. Maryam, M. Nadeem, and F. Anwer, "Helping Hearing-Impaired in Emergency Situations: A Deep Learning-Based Approach".

[2] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," CVPR, 2016.