Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation

Index

Introduction

- Communication is crucial, especially for the speech-impaired and hard-of-hearing communities.
- Importance of Sign Language Recognition (SLR)
- Sign languages are visual-based forms of communication and differ across cultures.
- Challenge: Bridging the communication gap between sign language users and non-users.

Research Problem

- Challenges in Sign Language Recognition
- Low recognition accuracy in current systems.
- Difficulties in real-time sign translation and video generation
- Misclassification, noise in existing systems.
- Lack of availability of Indian sign languages.

Comparison of existing SL recognition frameworks

Author	Technique	Static/ Dynamic	Category	Sign Language	Isolated/ Continuous	Accuracy
Barbhuiya et al. 2021 [36]	CNN+SVM	Static	Alphabets and Numbers	ASL	Isolated	99.82%
Aly et al. 2020 [37]	DeepLabv3+Bi-LSTM	Static	words	Arabic SL	Isolated	89.59%
Lee et al. 2020 [38]	LSTM+KNN	Static and Dynamic	Alphabets	ASL	Isolated	99.44%
Xiao et al. 2020 [39]	CNN+Bi-LSTM with attention	Static	Words, Sentences	CSL,GSL	Continuous	81.22% (CSL) 76.12% (GSL)
Elakkiya et al. 2021 [40]	GAN+LSTM+3DCNN	Dynamic	Sentences	GSL, ASL	Continuous	98.33%

Contribution of our project

- A hybrid CNN-BiLSTM model for recognizing signs and generating text.

- A system for translating spoken language into sign language using Neural Machine Translation (NMT).

- Generative Adversarial Networks (GAN) is used for generating high quality SL videos.

- Benchmarking with multilingual sign language datasets.

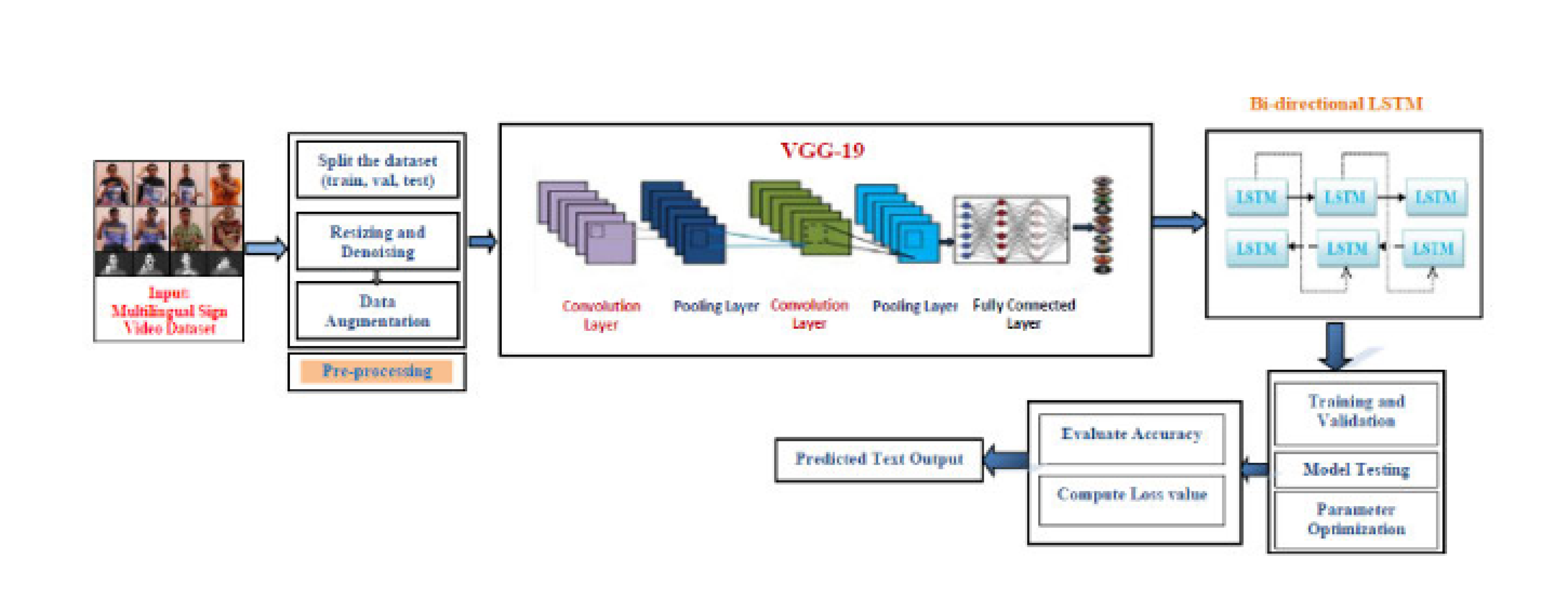
Related Work

- Previous Efforts in Sign Language Recognition
- Existing methods (CNN+SVM, Bi-LSTM) in different languages (ASL, Arabic SL, Chinese SL).
- Limitations in continuous sign language recognition.
- Need for sensor-less, real-time capable systems.

System Architecture Overview

- Hybrid Deep Neural Architecture (H-DNA)
- Hybrid CNN-BiLSTM for sign language recognition.
- MediaPipe for pose extraction.
- GANs for video generation.

High-level architecture of the proposed framework



Datasets

- RWTH-PHOENIX-Weather 2014T : German Sign Language, 40,000 videos.
- ISL-CSLTR:Indian Sign Language, 700 videos.
- How2Sign:American Sign Language, 2,456 videos.

Sign Language Recognition Model

CNN for Gesture Identification:

- VGG-19 model: 16 convolution layers, 3 fully connected layers
- Input images resized, processed through convolution and pooling layers
- Focuses on important features, reducing image size
- Outputs feature vector passed to Bi-LSTM

Bi-LSTM for Sequence Prediction:

- LSTM: Handles sequences using three gates (input, forget, output)
- Bi-LSTM: Uses past and future info for accurate gesture sequence prediction

Translation of Multilingual Sentences

- Nueral Machine Translation(NMT) and Attention Mechanism are used to convert spoken sentences into sign language.
- Recurrent Neural Networks (RNNs) to handle long sequences in language translation.

Translation Process & Results

- Word Embedding
- GRU Functions
- Attention Mechanism
- Output Generation
- Results: Evaluated on three major sign language datasets, showing strong performance

Pose Estimation

- MediaPipe provides human pose estimation for images and videos.

- Applications:
- Human activity tracking, sign gesture recognition, fraud monitoring, yoga pose analysis.

- Pose Estimation: Tracks poses of signers.

- Key Point Extraction: Extracts pose key points for further analysis.

- Deep Generative Networks: Uses key points to generate new poses.

Sign Video Generation

- Generated videos maintain realistic movements and smooth transitions.
- Visual results from the ISL-CSLTR and RWTH-PHOENIX datasets show high-quality, photo-realistic videos.
- Dynamic GAN outperforms earlier models in video generation for SL.

Conclusion

- Deep learning H-DNA framework for sign language recognition, translation, and video generation.
- Significant improvements over earlier

Key Metrics:

- 95%+ classification accuracy for SL recognition.
- 38.56 average BLEU score for translation.
- 3.46 FID2vid for video quality.
- 0.921 SSIM for structural similarity.
- 0.715 TCM for temporal consistency.

References

1. "Deep Learning-Based Standard Sign Language Discrimination"

Author Information: The authors of the paper are Menglin Zhang, Shuying Yang (corresponding author), Min Zhao

Objective: The system is designed to recognize sign language actions, distinguishing between correct (standardized) and incorrect (non-standardized) sign language actions. This is useful for improving sign language education software.

Accuracy: The model showed high accuracy, with the hand detection achieving up to 99.0% (mAP50) and the correctness discrimination model achieving about 81.6% accuracy for recognizing standard sign language actions using hand patches.

Disadvantages:

- High computational cost

2. "Sign Language Recognition: A Comprehensive Review of Traditional and Deep Learning Approaches, Datasets, and Challenges"

Author Information: The authors of the paper are Tangfei Tao, Yizhe Zhao, Tianyu Liu, and Jieli Zhu* from Xi’an Jiaotong University, China.

Objective: The review discusses various techniques for sign language recognition (SLR), focusing on both traditional methods and deep learning approaches. It highlights the challenges and advancements in feature extraction, temporal modeling, and dataset use.

Accuracy: For smaller datasets like CSL, accuracy can be very high. However, for larger datasets like WLASL, the accuracy drops significantly, reaching about 57.13%

Disadvantages:

- Real-World Application
- Generalization

3. “Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects”

Author Information: Bashaer A. Al Abdullah, Ghada A. Amoudi, and Hanan S. Alghamdi (King Abdulaziz University)

Objective: This document reviews advancements in sign language recognition (SLR) systems, emphasizing the use of AI and machine learning techniques, particularly deep learning models like CNNs, RNNs, and hybrid CNN-RNNs, to enhance recognition accuracy..

Author: The document reports that some DL models can achieve accuracy as high as 99% for certain tasks, but continuous recognition of sign language sentences remains a challenge.

Disadvantages:

- High computational cost
- Dataset limitations

4. "Hand Gesture Recognition for Multi-Culture Sign Language Using Graph and General Deep Learning Network"

Author Information: The authors of the paper are Abu Saleh Musa Miah, Md. Al Mehedi Hasan, Yoichi Tomioka, Jungpil Shin (Corresponding Author).

Objective: The system helps recognize hand gestures to translate different cultural sign languages (Korean, American, Japanese, etc.) into text.

Accuracy: It achieved 100% accuracy on the KSL-20 dataset and 99.60% on the ASL-20 dataset.

Disadvantages:

- Real-time use could be challenging due to the model's complexity.
- May struggle with recognizing new or unseen sign languages.

5. “Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective”

Author Information: The authors include notable figures like Danielle Bragg (Microsoft Research), Oscar Koller (Microsoft), and Hernisa Kacorri (University of Maryland).

Objective: The paper discusses the recognition, generation, and translation of sign languages. It emphasizes the need for interdisciplinary approaches combining computer vision, linguistics, natural language processing (NLP), and human-computer interaction.

Accuracy: The system relies on advanced AI techniques, but due to the lack of large, diverse datasets and the complexity of sign language, its accuracy is currently limited in real-world scenarios.

Disadvantages:

- Limited Dataset
- Lack of Real-world Applications

Thank You!