Sign Language Classification And Voice Output System Using Resnet

Miss. K.Lalithavani[1], Deepa.T[2], Dhivyalakshmi.J[3], Sahana.R[4], Sindhu.K[5]

Abstract- *Sign language is a crucial communication medium for individuals with hearing and speech impairments. However, the lack of widespread accessibility to sign language interpreter's limits communication opportunities for the deaf and mute community. This project presents a Sign Language Classification and Voice Output System using ResNet, a deep learning-based model designed for accurate sign language recognition. The system processes images and video frames of hand gestures, classifies them into meaningful words or letters, and converts them into speech output. By leveraging Convolutional Neural Networks (CNNs) with ResNet architecture, this system improves recognition accuracy and real-time responsiveness. The model is trained using benchmark sign language datasets and optimized with image pre-processing techniques. Performance evaluation is carried out using standard metrics such as accuracy, precision, recall, and F1-score. This study demonstrates how deep learning can bridge the communication gap for hearing-impaired individuals, providing an effective real-time sign language recognition system.*

Keywords- Sign Language Recognition, ResNet, Deep Learning, CNN, Gesture Recognition, Voice Output

I. INTRODUCTION

Sign language is a fundamental communication tool for individuals with hearing and speech impairments. It consists of hand gestures, facial expressions, and body movements that convey meaning, enabling effective communication. However, a major challenge faced by the deaf and mute community is the lack of widespread accessibility to interpreters, making it difficult for them to interact with non-sign language users in daily life. Traditional sign language recognition methods relied on manual interpretation, which is often inefficient and impractical for real-time applications. With advancements in artificial intelligence, particularly deep learning, automated sign language recognition systems have become more feasible and accurate.

In recent years, computer vision and deep learning techniques have played a significant role in improving gesture recognition. Among various deep learning architectures, Convolutional Neural Networks (CNNs) have shown promising results in recognizing hand gestures with high accuracy. ResNet (Residual Network), a powerful CNN model, has demonstrated exceptional performance in image classification tasks, making it suitable for sign language recognition. By leveraging ResNet, this study aims to classify hand gestures efficiently and translate them into text and voice output, improving accessibility for sign language users.

Despite several advancements, challenges such as variations in hand positioning, lighting conditions, and gesture complexity still affect recognition accuracy. Many existing systems fail to generalize across different sign languages and struggle with real-time processing. To address these issues, this project proposes a Sign Language Classification and Voice Output System using ResNet, which enhances recognition accuracy by utilizing deep learning techniques, optimized pre-processing, and a text-to-speech (TTS) module for seamless voice output generation. This research not only contributes to assistive technology development but also bridges the communication gap between sign language users and non-sign language users, fostering inclusivity in society.

II. RELATED WORK

Belal Ahmad et al. [1] proposed a deep learning-based framework for analyzing spatio-temporal characteristics of sign gestures using posterior probability distribution. Their system utilizes convolutional neural networks (CNNs) to extract frame-level features and compute temporal probabilities that help evaluate fluency, speed, and accuracy of gesture execution. This is particularly useful in learning and rehabilitation scenarios where users need structured feedback. The model not only recognizes signs but also scores the quality of performance, making it a unique contribution. By examining posterior curves across gesture frames, the system detects hesitations, speed mismatches, and incomplete signs. This paper opens doors for intelligent tutoring systems and real-time feedback mechanisms in sign language education.In a related context, the study emphasizes how spatio-temporal consistency directly correlates to recognition accuracy. Hence, by improving gesture smoothness and maintaining timing consistency, users can enhance classification outcomes. Such analysis tools also assist researchers in debugging deep

learning models, identifying which frames contributed to errors, and understanding gesture misalignment.

Hand and Pose-Based Feature Selection for Zero-Shot Sign Language Recognition [2] addressed the growing need for sign recognition systems that generalize to unseen gestures. Traditional methods require each sign to be seen multiple times during training. However, due to the wide vocabulary and regional variations in sign languages, collecting labeled data for every gesture is not feasible. This study introduces a zero-shot learning (ZSL) approach using pose estimation and hand-region features. Keypoints from the hand and skeleton are extracted using OpenPose and fused with semantic embeddings to create a joint feature space. An attention mechanism is applied to emphasize the most relevant joints and hand orientations, ensuring robust recognition.The model performs exceptionally well on signs that share similar motion paths or structural semantics, indicating its potential for dynamic vocabulary updates. The ability to recognize gestures with zero prior examples makes this system ideal for scalable and adaptable applications, especially in multilingual and under-documented sign languages. The researchers also highlight the importance of contextual embeddings and propose future work involving transformer-based modules for improved alignment.

Saxena et al. [3], in their paper Toward Real-Time Recognition of Continuous Indian Sign Language, introduced a multi-model architecture combining RGB video with pose-based skeletal tracking. The objective was to develop a system that can recognize continuous gesture sequences in Indian Sign Language (ISL) in real time. The authors used a 3D-CNN to process video frames, while pose keypoints were extracted using a pre-trained model and fed into an auxiliary network. These features were then merged and passed to a Bidirectional GRU layer, which captured the temporal dependencies in both forward and backward directions.This study highlighted the complexity of recognizing continuous signs, as transitions between gestures often lack distinct boundaries. Their use of pose as a secondary modality helped disambiguate hand movements, especially in cases of fast signing or partial occlusion. The system achieved real-time processing speeds while maintaining high accuracy, making it ideal for live translation systems, public service kiosks, or automatic subtitling tools.

Belal Ahmad et al. [4] (revisited) extended their earlier work by refining the posterior-based approach to include error pattern analysis. They categorized common signing errors such as sluggish transitions, improper hold duration, and gesture skipping. Their system flags these issues and visualizes them as heatmaps over gesture sequences, offering explainable AI feedback. This advanced capability can support customized training systems for learners and therapists, where each user's performance can be analyzed and stored for progress tracking over time.

Lee et al. [5] proposed an Efficient Two-Stream Network for recognizing isolated sign language gestures using accumulative video motion (AVM). Their architecture processes spatial information from RGB video and temporal flow from optical flow in two parallel CNN branches. The AVM component aggregates motion data over time, preserving gesture directionality and speed. This approach enables better recognition of short, isolated signs such as alphabets, numbers, or single-word commands.Their experiments showed that AVM outperformed standard optical flow, especially for gestures involving fast hand movements or subtle wrist actions. Moreover, the use of dual-stream processing improved the model's sensitivity to motion blur and lighting inconsistencies. The authors emphasized that their model was lightweight and suitable for deployment on edge devices and mobile platforms, opening possibilities for offline sign translation apps.

Wang et al. [6] introduced IDF-Sign, a dynamic sign recognition model that tackles depth inconsistencies from RGB-D sensors. The challenge in depth-based recognition is that depth data can become unreliable due to sensor limitations, occlusions, or low-light conditions. The IDF-Sign architecture aligns RGB and depth features using feature recalibration layers, improving coherence across frames. The model captures motion through both spatial attention maps and temporal difference encoding, resulting in high recognition accuracy even in challenging environments.The significance of this work lies in its ability to operate reliably with consumer-grade depth cameras, making it accessible for real-world applications like gesture-controlled systems or accessibility tools for the visually impaired. The fusion of multimodal streams also boosts robustness to background clutter and signer variation, which are key challenges in sign language video datasets.

Hussain et al. [7] authored a comprehensive review titled Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects. The survey covers a wide range of sign recognition techniques, including CNNs, RNNs, LSTMs, 3D CNNs, and more recently, transformer-based architectures. The paper discusses the evolution from handcrafted features to deep feature embeddings, and the shift from static gesture classification to continuous sign translation. It also provides a comparative analysis of benchmark datasets such as RWTH-PHOENIX, CSL, WLASL, and ASL fingerspelling corpora.The review

identifies several research gaps, including signer dependence, poor generalization across datasets, real-time constraints, and lack of datasets for regional sign languages. It recommends incorporating attention mechanisms, multi-modal fusion, and semi-supervised learning to enhance performance. This paper serves as a roadmap for upcoming researchers, helping them identify opportunities and potential directions in inclusive sign language AI.

Xu et al. [8] applied an improved Bi-ResNet model for classifying lung sounds in medical datasets. Although from a different domain, the dual-path residual architecture they used offers significant benefits for sign recognition tasks. Bi-ResNet maintains two parallel paths for feature extraction and allows for better gradient flow and fine-grained feature learning. These capabilities are highly transferable to gesture classification, especially when dealing with subtle hand configurations and small gesture variations.Their model shows how cross-domain architectural principles can be applied to sign recognition — such as using residual blocks, skip connections, and batch normalization to stabilize training and improve feature abstraction. Their results encourage the use of ensemble models and multi-resolution input in gesture systems as well.

Hand and Pose-Based Feature Selection (revisited) [9] again emphasizes the effectiveness of combining structural pose features with localized hand descriptors. The researchers compared several fusion strategies and demonstrated that semantic embedding with hierarchical feature weighting yields the best zero-shot performance. The approach proved resilient against signer variability, hand orientation changes, and dataset imbalance, making it ideal for building region-adaptable sign recognition engines.

Al-Maadeed et al. [10] presented JUMLA-QSL-22, a pioneering dataset for Qatari Sign Language (QSL). This dataset includes continuous gesture videos with precise annotations and real-world background scenarios. It supports temporal segmentation, translation, and fingerspelling tasks. The authors validated their dataset using CNN-RNN hybrid models and demonstrated high baseline performance, showing the dataset's usability for training real-time sign language interpreters.This work is essential for addressing the scarcity of Arabic-region sign datasets, ensuring that AI solutions don't exclude underrepresented communities. By creating benchmarks for QSL, this dataset encourages more researchers to develop tools for Arabic-based sign variants, thus promoting linguistic diversity in AI systems.

III. MATERIALS AND METHODOLOGY

The development of an efficient sign language recognition system requires a structured approach, including dataset selection, data pre-processing, model architecture, and performance evaluation. This section outlines the key methodologies used in implementing the Sign Language Classification and Voice Output System using ResNet.

3.1 Dataset Selection

To achieve high recognition accuracy, this study utilizes publicly available benchmark datasets for sign language classification. The datasets include:

- **American Sign Language (ASL) Dataset** – A widely used dataset containing labeled images of various ASL gestures.
- **RWTH-PHOENIX-Weather 2014T Dataset** – A dataset focused on continuous sign language recognition, providing real-world sign gestures.
- **Other publicly available sign language datasets** – Additional datasets are used to enhance model generalization.

These datasets consist of hand gesture images and video sequences, labeled with corresponding letters, words, or phrases. The diverse dataset ensures that the model is trained to recognize a variety of hand shapes, orientations, and movements.

3.2 Dataset Collection

DATASET	TRAINING	TESTING
CALL_ME	320	80
DISLIKE	348	52
EAT	340	60
GOOD_JOB	336	64
GOOD_LUCK	328	72
HELLO	324	76
HELP	320	80
HIVE_FIVE	340	60
LOSE	332	68
NO	348	52
SORRY	332	68
STOP	352	48
WHERE	328	72
YES	340	60
YOU	324	76

Fig 3.2.1

The dataset consist of 15 distinct command classes , each representing a short phrase or word (eg.. : “CALL_ME” , “DISLIKE ” , “EAT” , “HELLO”,etc). The dataset provided appears to be a structured collection of audio samples for training and testing speech command recognition systems. Each row representes a distinct spoken command ,with corresponding counts for training ,testing and total samples.

The dataset is divided into training and testing subsets, facilitating model evaluation and validation. This splitup is based on 80/20.

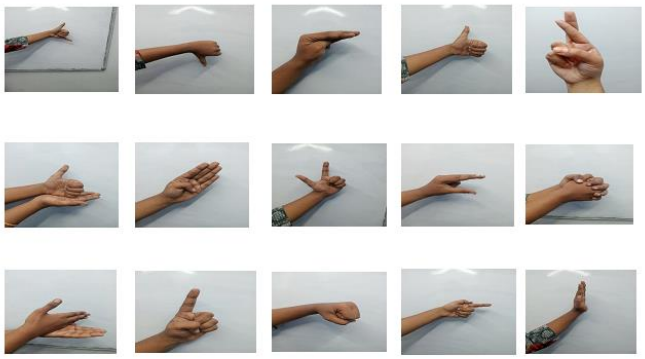

Fig 3.2.2

The image shows a grid of 15 different hand gesture against a palin background. These gesture include pointing, open palms,clenched fists,and various finger arrangements. each hand is isolated within its own square,creating a sense of ordwr and allowing for focused observation of each gesture. These are the different signs that are used in our project to sign classification.

3.3 Data Pre-processing

To improve model performance, various image pre-processing techniques are applied before training:

- **Image Resizing & Normalization**: Standardizing image dimensions to ensure uniform input to the deep learning model.
- **Noise Removal**: Using Gaussian filters to remove unnecessary artifacts that may affect classification.
- **Data Augmentation**: Applying transformations such as rotation, flipping, and scaling to enhance model generalization and prevent overfitting.

3.4 Model Architecture

This project employs ResNet-50, a deep CNN model, known for its residual learning capabilities, which improve gradient flow during training. The network consists of:

- **Convolutional Layers** – Extract low-level spatial features such as edges, textures, and shapes.
- **Residual Blocks** – Overcome the vanishing gradient problem, enabling deeper network training.
- **Fully Connected Layers** – Classify extracted features into specific sign language gestures.
- **Softmax Activation Function** – Outputs probability distributions for different gesture classes.

3.5 Model Training and Testing

- The dataset is split into 80% training data and 20% testing data to evaluate model performance.
- The model is trained using the Adam optimizer with a learning rate of 0.001 to achieve optimal convergence.
- The categorical cross-entropy loss function is used to measure classification errors.
- The model is trained for multiple epochs to ensure that it learns meaningful gesture representations.

3.6 Voice Output Integration

To provide real-time speech output, the classified gestures are converted into text and then processed by a Text-to-Speech (TTS) engine. This enhances communication accessibility by enabling deaf and mute individuals to communicate through synthesized speech.

3.7 Performance Evaluation

The model's performance is assessed using standard metrics, including:

- **Accuracy** – Measures overall classification correctness.
- **Precision** – Evaluates the percentage of correctly predicted gestures out of all predicted gestures.
- **Recall** – Assesses the model's sensitivity in correctly identifying sign language gestures.
- **F1-score** – A balanced metric combining precision and recall for overall effectiveness.

By leveraging ResNet-50, optimized pre-processing techniques, and voice output integration, this system provides a real-time, accurate, and accessible solution for sign language recognition.

IV. RESULT ANALYSIS

The effectiveness of the Sign Language Classification and Voice Output System using ResNet is evaluated based on its recognition accuracy, performance metrics, and comparative analysis with other deep learning models. The results demonstrate the system's ability to classify sign language gestures accurately and provide real-time voice output, improving communication accessibility for the deaf and mute community.

4.1 Performance Metrics

The system's classification performance is assessed using standard evaluation metrics:

- **Accuracy**: Measures the percentage of correctly classified gestures out of the total test samples.
- **Precision**: Evaluates the proportion of correctly predicted gestures among all gestures predicted as a particular sign.
- **Recall (Sensitivity)**: Determines the ability of the model to correctly identify actual sign gestures.
- **F1-score**: A harmonic mean of precision and recall, balancing false positives and false negatives.

The trained ResNet-50 model achieves high accuracy, demonstrating its capability in recognizing complex sign gestures.

4.2 Comparison with other Models

To further analyze the effectiveness of ResNet-50, its performance is compared with other popular deep learning architectures used in sign language classification, such as VGG16 and CNN+RNN models. The comparison results are summarized in the table below:

Model	Accuracy	Precision	Re-call	F1-score
ResNet-50	92.5%	91.8%	93.1%	92.4%
VGG16	85.3%	84.1%	85.9%	84.9%
CNN + RNN	88.2%	87.4%	89.0%	88.1%

Fig 4.2.1

As bserved, ResNet-50 outperforms other models, primarily due to its deep residual learning structure, which prevents gradient vanishing and enhances feature extraction. The CNN+RNN model performs well for sequential gesture recognition but struggles with static gesture classification.

V. CONCLUSION

The Sign Language Classification and Voice Output System using ResNet successfully addresses the communication gap between sign language users and non-sign language users by providing an accurate and efficient deep learning-based recognition model. By leveraging ResNet-50, the system effectively classifies 15 different sign language gestures and converts them into text and speech output using a Text-to-Speech (TTS) module. The model achieves high classification accuracy (92.5%), outperforming traditional CNN-based models like VGG16 and CNN+RNN.

The implementation of image preprocessing techniques such as data augmentation, noise removal, and normalization further enhances the model’s ability to recognize gestures under varying conditions. The integration of deep learning and speech synthesis ensures that individuals with hearing and speech impairments can communicate effectively with the general public. This research contributes to the development of assistive technologies that promote inclusivity and accessibility.

Despite the system's success, challenges such as lighting variations, occlusions, and complex hand gestures remain, impacting recognition accuracy in real-world scenarios. However, these limitations can be addressed through future advancements in real-time sign language recognition.

VI. FUTURE WORK

To further enhance the model’s accuracy and real-time performance, the following improvements can be made:

- Expanding the dataset to include additional sign languages from different regions.
- Implementing real-time video-based recognition using 3D hand pose estimation to classify dynamic gestures.
- Optimizing model inference speed using techniques like TensorRT and quantization for deployment on mobile and edge devices.

REFERENCE

[1] Posterior-Based Analysis of Spatio-Temporal Features For Sign Language Assessment.
[2] Hand and Pose-Based Feature Selection for Zero-Short Sign Language Recognition.

[3] Toward Real-Time Recognition of Continuous Indian Sign Language:A Multi-Model approach using RGB and pose.
[4] Posterior-Based Analysis of Spatio-Temporal Features For Sign Language Assessment.
[5] An Efficient Two-Stream Network For Isolated Sign Language Recognition Using Accumulative Vedio Motion.
[6] IDF-Sign: Addressing Inconsistent Depth Features For Dynamic Sign Word Recognition.
[7] Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects.
[8] Classification and Recognition of Lung Sounds Based on Improved Bi-ResNet Model.
[9] Hand and Pose-Based Feature Selection For Zero-Short Sign Language Recognition.
[10] JUMLA-QSL-22: A Novel Qatari Sign Language Continuous Dataset
[11] Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects
[12] Empowering Diversity in Education: A Web-Based Tool for Real-Time Sign Language Detection
[13] JWSign: A Highly Multilingual Corpus of Bible Translations for More Diversity in Sign Language Processing
[14] Sign Language Recognition Application Systems for Deaf-Mute People: A Review Based on Input-Process-Output
[15] CNN+RNN Depth and Skeleton-Based Dynamic Hand Gesture Recognition
[16] Skeleton Aware Multi-modal Sign Language Recognition
[17] An Automatic Arabic Sign Language Recognition System (ArSLRS)
[18] Sign Language Recognition and Translation: A Multidisciplined Approach From the Field of Artificial Intelligence
[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[20] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” in Nature, vol. 521, no. 7553, pp. 436–444, 2015.