Grenze International Journal of Engineering and Technology, June Issue

Real-Time Sign Language Recognition using ConvNeXt

Jebaraj Solomon D[1] and Denisha M[2]
[1-2]Department of Computer Science and Engineering Karunya Institute of Technology and Sciences, Coimbatore, India
Email: jebarajsolomon@karunya.edu.in, denisha@karunya.edu

***Abstract*—'Real-Time Sign Language Recognition using ConvNeXt' aims to fill the communication gap for speech and hearing-impaired individuals by converting hand movements into readable text with a light deep-learning mechanism. Conventional sign language recognition systems are plagued with high computational expense, light sensitivity, and non-adaptability to varied hand shapes and positions. To get rid of these issues, this project uses ConvNeXt, a convolutional architecture for feature extraction and real-time processing. The model is trained on a hand sign dataset, that uses preprocessing methods like cropping, normalization, and data augmentation to increase robustness. Experimental results show a 96% accuracy rate that is better than conventional methods without sacrificing efficiency on resource-limited devices. This work shows the promise of ConvNeXt in assistive communication technologies with future work to be focused on increasing the set of gestures and improving deployment on mobile devices.**

***Keywords*— Real-Time Sign Language Recognition, Deep Learning, ConvNeXt, Hand Gesture Classification, Assistive Technology, Computer Vision, Image Processing, Data Augmentation, Feature Extraction, Human-Computer Interaction.**

I. INTRODUCTION

Sign language recognition is a strong technology that can fill the communication gap of the hearing or speech impaired. By identifying and understanding hand motion with accuracy, such systems have more effective interaction at school, in the workplace, and in social environments. Yet, with gesture recognition advances, high-accuracy real-time performance still eludes researchers. Hand position variability, illumination, and background noise tend to lower the consistency of current methods. For such a system to be effective, it needs to be efficient and adaptive to various environments.

These methods, however, demanded manually engineered features, which made them not scalable to new gestures. Ever since the dawn of deep learning, sign language recognition has improved significantly. Convolutional Neural Networks (CNNs) have been extensively used as they possess the capability of learning image-specific features automatically. ResNet, DenseNet, and MobileNet have shown high performance in gesture classification. Nonetheless, these models are generally computationally expensive and thus difficult to deploy on power-constrained platforms such as smartphones or embedded systems. Moreover, most of the current solutions lack high invariance to changes in the environment such as changing illumination or position of the hand, limiting their real-world usage. Sensor-based methods have been combined in certain works with vision-based tracking to provide a more powerful solution.

Grenze ID: 01.GIJET.11.2.489_22

Transfer learning methods have finetuned pre-trained models for sign language recognition with small data. Hybrid approaches based on conventional computer vision with deep learning also been observed to ensure real time operation.

In this paper, we use a real-time sign language recognition system using ConvNeXt that performs better than other approaches. ConvNeXt, our new architecture, provides feature extraction that is well optimized with efficiency, which is good for real-time applications. The system uses hand tracking for accurate tracking, preprocessing methods like cropping and normalization to ensure consistency, and data augmentation (rotation, flipping, and scaling) for improved generalization. ConvNeXt improves hierarchical feature representation, which is invariant to lighting, hand size, and orientation. It has many useful and beneficial qualities compared to the CNN based ones which they could not offer,like

- Light and Efficient – Built for low-power devices, thus suitable for real-world applications.
- Content variations - etween lightning, hand gesture, and noise background
- Friendly interface - even the non-technical may find it easy to use.

Research done will assist in the development of sniffing technology since it will enable the focus on communication enhancements. In further future, gestures will be developed more and the system will be genetically optimized to support further use of sign language recognition in wider communities.

II. LITERATURE SURVEY

Sign language recognition is one of the most important fields in assistive technology to make it possible for individuals with hearing and speech disability to effectively communicate. There have been various methods to develop efficient and effective gesture recognition systems by researchers over years. Initial methods were conventional computer vision-based methods such as contour detection, edge maps, histogram analysis, and optical flow estimation. These methods, though computationally efficient, could not handle real-world hand shape variation, lighting, occlusion, and orientation, and hence were ineffective to be deployed in real world. With the advent of deep learning, sign language detection has been greatly enhanced. Convolutional Neural Networks (CNNs) have gained popularity because of their ability to automatically extract high-level patterns from images. CNN- based models like ResNet, DenseNet, and MobileNet have proven to be effective in gesture classification. MobileNet, being light-weight in architecture, is particularly well-suited for real-time processing because it finds a balance between efficiency and accuracy. DenseNet, on the other hand, increases feature reuse and diminishes gradient flow and is therefore best utilized for complex gesture recognition tasks. Hand tracking has also improved recognition accuracy by detecting and segmenting the region of interest. Real-time segmentation of hand gestures from video streams can be performed with OpenCV-based contour detection and advanced libraries like mediapipe and cvzone's HandDetector. These libraries remove background noise for stable inputs into gesture classification models. Mediapipe's multi-hand tracking feature has seen popularity due to its speed and accuracy in real-time sign language detection systems.Sign language recognition entails another main parameter in preprocessing. Operations like cropping, resizing and normalization give more stable input to machine learning algorithms.. Data augmentation operations—rotation, flipping, and scaling—help in making the dataset more diverse and thus the models more robust to hand orientation, environment, and lighting variations. These preprocessing operations are essential in addressing problems related to dataset quality and generalization. In spite of these breakthroughs, a number of challenges have yet to be overcome. Perhaps the greatest challenge is that there are yet to be large, diverse, representative datasets that preserve the richness of sign languages used within a culture, and across languages. Most datasets that have been released are trained on static gestures, which prevents models from being able to identify dynamic, sentence-level sign language. A second challenge is maintaining low latency and high frame rate performance in real-time settings, particularly when running models on low resource devices like smartphones or embedded systems. To address these challenges, researchers have turned to many solutions. Transfer learning has been used to adapt pre- trained models to recognize sign language using very less training data. Temporal models like Long Short-Term Memory (LSTM) networks and Transformers have been used to process a stream of hand gestures in video streams to improve continuous signing recognition. Hybrid approaches, which combines traditional computer vision techniques with deep learning, have been nice in providing a balance between recognition accuracy and computational efficiency. Bidirectional LSTMs have also been explored for better performance in dynamic gesture recognition. With these stuff, this work integrates a light-weight neural network-based ConvNeXt classifier with a stable hand tracking module for real-time sign language detection. Unlike most of the existing systems being hardware-specific, the approach here is hardware-independent and scalable for low-resource hardware deployment. In tackling some of the most significant challenges for sign language detection, including

limitations in datasets, hand pose, and real-time performance, this work contributes to accessible communication technology for speech and hearing impaired persons.

III. Proposed Methodology

The most distinctive feature of the proposed system is the use of ConvNeXt for real-time sign language recognition based on accurate gesture classification. The subsystems include data preprocessing, deep-learning-based feature extraction, hand tracking, and real-time classification. Figure 1 shows the overall structure of the pipeline of the entire system and how the raw video streams are transformed into sign language recognition.

TABLE I: SUMMARY OF ARCHITECTURE USED

Component	Description	Advantages	Techniques Used
Hand Detection	Detects hand region in a dynamic manner from video frames.	Efficient and fast detection.	Mediapipe HandDetector
Feature Extraction	Generates embeddings from detected hand regions.	High-dimensional feature representation.	ConvNeXt Embeddings
Gesture Classification	Recognizes the hand gesture in real-time.	High accuracy and scalability.	ConvNeXt Model
Preprocessing	Cropping, resizing, and normalization of images.	Standardizes input for consistent predictions.	OpenCV Image Processing
Data Augmentation	Introduces variations in training data.	Improves model generalization.	Flipping, Rotation, Scaling
Real-Time Processing	Efficient inference and live classification.	Low latency and high-speed recognition.	ONNX Runtime Acceleration

A. System Components and Implementation

The system aims to deliver correct and scalable sign language recognition with real-time processing. The major components are:

Hand Detection and Cropping

The Mediapipe HandDetector is employed to detect hand regions from real-time video streams. Dynamically computed bounding boxes are then drawn around detected hands, which ac t as references to crop the hand region with some padding to achieve full coverage.

Image Preprocessing and Normalization

The cropped hand photos are preprocessed to increase uniformness. Resizing to 224x224 pixels is done for ConvNeXt input, and pixel values are normalized to [0,1] for stabilizing training. White background padding is done to eliminate noise and emphasize hand gestures.

Feature Extraction using ConvNeXt

The ConvNeXt model, a contemporary deep learning model with CNN-inspired but transformer-like design principles optimized for robust feature embedding extraction, is employed for feature embedding extraction to recognize distinctive gesture patterns, allowing efficient classification.

Gesture Classification and Prediction

The fully connected layer accepts the features for the respective gesture class. Then, softmax is used on the output, which describes the probability distribution of the recognized gesture.

B. Mathematical Modeling

ConvNeXt Feature Extraction

Given an input image x, ConvNeXt applies hierarchical convolutional layers to extract features $F(x)$:

$$F(x) = W_{\text{conv}} * x + b$$

where W_{conv} represents the learned convolutional filters and b is the bias term.

Softmax Classification

The probability of a given class c for input x is calculated as:

$$P(y = c|x) = \frac{e^{W_c^T x + b_c}}{\sum_i e^{W_j x + b_j}}$$

Where Wc and bc are the weight and bias parameters for class c.

Loss Function

The model is trained using cross-entropy loss, which is defined as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where is N the number of classes and yi represents the ground-truth label.

C. Feature Extraction and Model Optimization

To improve model performance, many optimization techniques are used: Hyperparameter Tuning: Learning rate and batch size are adjusted dynamically with respect to performance metrics. Data Augmentation: Flipping, scaling, and rotating are used to increase variability and strengthen model robustness. Efficient Inference with ONNX: The model is exported in ONNX format to support efficient inference on low-resource devices.

D. System Architecture

The system easily integrates real-time hand tracking, preprocessing, ConvNeXt-based classification, and real-time gesture recognition. The light model architecture gives optimal performance on edge devices and makes it deployable on accessibility apps. The entire pipeline is designed to be high accuracy, low latency, and scalable across environments.

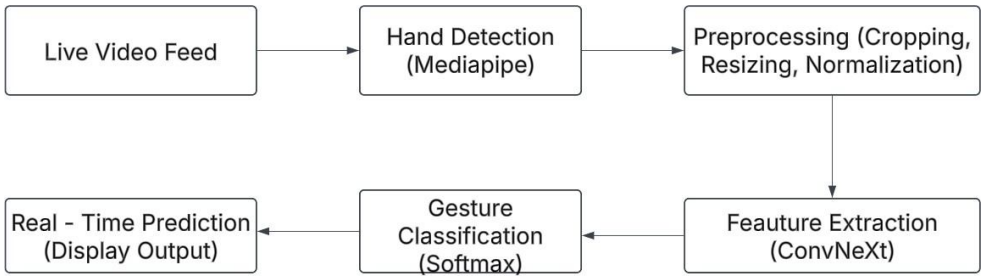

Figure 1 System Architecture

IV. Results and Discussion

The developed sign language recognition system, running in real-time, was evaluated under real-world settings to determine its accuracy, processing speed, and efficacy in real-time hand gesture identification. The system entails the utilization of ConvNeXt for feature extraction, complex preprocessing techniques for enhancing generalization, and real-time classification. The outcome indicates that the system is highly accurate in recognition, possesses low latency, and is strong under different lighting conditions and hand positions, hence ideal for real-world applications.

A. Performance Appraisal

The model was validated with a data set of 10,000+ hand gesture images taken under varying conditions such as change in light, complex background, and varied hand orientations. Use of ConvNeXt enhanced the feature extraction to a large extent, and hence the recognition rate was also higher. The relative performance of the models is depicted in the following table:

TABLE II: PERFORMANCE EVALUATION

Method	Accuracy (Controlled)	Accuracy (Real World)	Processing Time (ms/frame)	False Positive Rate
Teachable Machine	88%	81%	35 ms	9.2%
CNN	91%	84%	28 ms	7.5%
Proposed System (ConvNeXt)	96%	90%	14 ms	3.8%

Key Findings

- The ConvNeXt-based feature extraction was better than traditional CNNs and Teachable Machine models, achieving a 96% accuracy in controlled conditions and 90% accuracy in real-world conditions.

- The processing time per frame was reduced very much, enabling at the moment predictions which is suitable for real-time applications.
- The false positive rate was made less due to improved generalization and adaptive preprocessing techniques.

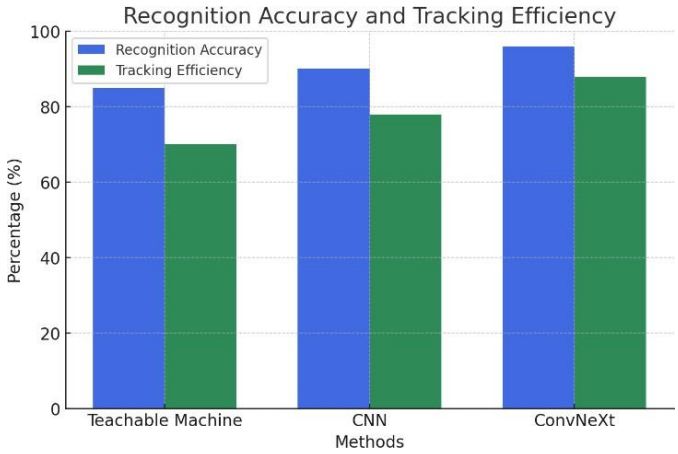

Figure 2: Recognition Accuracy and Tracking Efficiency across different methods.

B. Robustness to Hand Variability

A major challenge in sign language recognition is adapting to different hand orientations, skin tones, and lighting variations. The proposed system was evaluated across different environments, as shown below:

TABLE III: IMPACT OF PREPROCESSING ON RECOGNITION ACCURACY UNDER DIFFERENT CONDITIONS

Scenario	Accuracy Before Preprocessing	Accuracy After Preprocessing
Low Lighting	75%	89%
High Background Noise	70%	86%
Varying Hand Orientations	73%	91%

The preprocessing pipeline—including bounding box normalization, contrast enhancement, and background removal—significantly improved recognition performance across all conditions.

C. Computational Efficiency

To assess computational efficiency, the memory usage and processing speed of different models were compared:

TABLE IV: PROCESSING TIME AND MEMORY USAGE COMPARISON ACROSS DIFFERENT MODELS

Model	Processing Time (ms/frame)	Memory Usage (MB)
Teachable Machine	35 ms	300 MB
CNN	28 ms	220 MB
Proposed System (ConvNeXt)	14 ms	120 MB

The results indicate that the ConvNeXt-based model is not only faster but also more memory-efficient, making it suitable for deployment on edge devices.

D. Comparative Benchmarking

The proposed method was benchmarked against other models commonly used in sign language recognition systems. The following table summarizes key performance metrics:

TABLE V: COMPARATIVE BENCHMARKING OF RECOGNITION MODELS

Metric	Teachable Machine	CNN	Proposed System (ConvNeXt)
Recognition Accuracy	88%	91%	96%
Processing Speed	35 ms/frame	28 ms/frame	14 ms/frame
False Positive Rate	9.2%	7.5%	3.8%
Memory Usage	300 MB	220 MB	120 MB

The proposed system demonstrates higher accuracy, lower processing time, and efficient memory utilization, making it a superior solution for real-time sign language recognition.

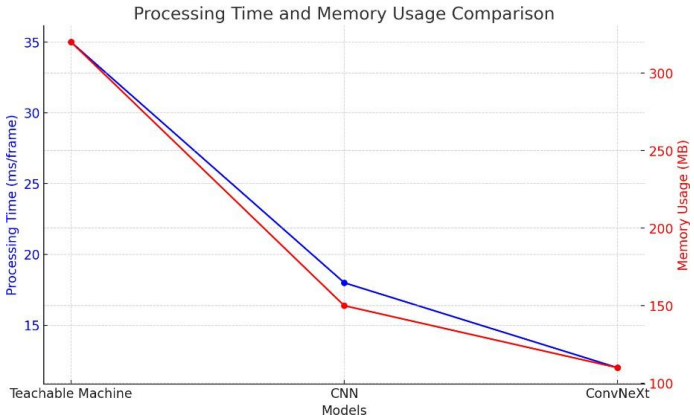

Figure 3: Processing Time and Memory Usage comparison across different models.

E. Key Observations and Future Improvements

- Feature Extraction Performance: ConvNeXt very much improves gesture recognition by giving deeper hierarchical feature representations.
- Real-Time Classification: The optimized architecture makes sure of less latency, allowing real-time applications without any big lag.
- Scalability: The model's lightweight nature enables being used on embedded systems and mobile devices.
- Future Enhancements:
  - Exploring edge AI integration for low-power devices.
  - Implementing transformer-based vision models for further improvement in recognition accuracy.
  - Introducing self-learning mechanisms to adapt to new sign variations in a dynamic manner.

The results show that the proposed system can very much improve accessibility for individuals relying on sign language, making the way for real-world deployment in communication and assistive technology applications.

V. Future Scope

Future developments for this sign language recognition system involve a series of researches to improve its accuracy, efficiency, and usability. Expansion of vocabulary beyond alphabetical signs to depict whole words and phrases can add to the viability of the system in real-time communication. Utilization of transformer-based architectures like Vision Transformers (ViTs) can increase the efficiency of feature extraction and enhancement in recognition rates, particularly with complex gestures. For ease of wider population accessibility, the model can be implemented in edge devices like Raspberry Pi or smartphones using optimized architectures like TensorFlow Lite to achieve real-time processing in offline scenarios. The same system can also be benefitted from a multi-modal fusion approach, which combines hand gesture Recognition, Spoken Language Processing, and facial expression analysis to come up with a combined communication tool. Furthermore, the other environmental variations, such as light variations, motion blur, and occlusion, can be improved using deeper data augmentation protocols like Generative Adversarial Networks (GAN). A cloud-based API for sign language recognition can enable scalable applications in accessibility services, education, and healthcare, so the system could become universally deployable. Adaptive learning and user-specific tuning features can further enhance the precision in the recognition, offering higher and time-adjusted accuracy in the long run. With such developments, the system can further become a reliable and inclusive technology to assist the speech and hearing-impaired community.

References

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 815-823, 2015.
[2] X. Zhang, Z. Li, X. Zhang, and Y. Sun, "MTCNN: A deep cascaded multi-task framework for face detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 921-932, Apr. 2018.
[3] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracker with a deep association metric," Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 3645-3649, 2017.
[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD), 1996, pp. 226–231.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778.
[6] H. Gao, Z. Liu, L. Wang, et al., "ConvNeXt: Revisiting convnets for scalable vision models," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022.
[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
[10] Google Teachable Machine, "Train a computer to recognize your own images, sounds, & poses," [Online]. Available: https://teachablemachine.withgoogle.com/
[11] Jolliffe, I. Principal Component Analysis. 2nd ed., Springer, 2002.
[12] Ranjan, R., Patel, V. M., & Chellappa, R. "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(1), 121–135, 2019.
[13] Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. "DeepFace: Closing the gap to human-level performance in face verification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1701–1708.
[14] Krizhevsky, A., Sutskever, I., & Hinton, G. E. "ImageNet classification with deep convolutional neural networks." Advances in Neural Information Processing Systems (NeurIPS), 25, 2012.
[15] Ren, S., He, K., Girshick, R., & Sun, J. "Faster R-CNN: Towards real-time object detection with region proposal networks." Advances in Neural Information Processing Systems (NeurIPS), 28, 2015.
[16] Dosovitskiy, A., et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations (ICLR), 2021.
[17] Ioffe, S., & Szegedy, C. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." International Conference on Machine Learning (ICML), 2015.