International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Dynamic Hand Gesture Recognition for Sign Language: Translating Sign Language through Gesture Patterns

Jayavelu Balaji

Submitted: 10/01/2024 **Revised:** 20/02/2024 **Accepted:** 27/02/2024

Abstract— Sign language is the key to information exchange in the Deaf and Hard of Hearing Communities. Nevertheless, the communication between signers and non-signers will be constrained if interpreters must fulfil the translation process from sign language to spoken or written language. Over the past few years, the flourishing computer vision and machine learning have come out in a sprouting of highly competitive hand gesture recognition systems based on sign language translation. This article introduces a new method for the recognition of dynamic hand movements for translating sign language into text dependent on the feature extraction and distinction of the gesture patterns through Convolutional Neural Networks (CNN). The proposed algorithm traces and encodes fast hand gestures in living moments. Features are retrieved, and thus, the gestures of sign language are correctly categorized and presented in text or speech-language format. The paper summarizes the training and testing of the neural networks, the algorithm outputting promising results in the aspect of accuracy and efficiency, finding its applicability in the expanding communication of the deaf and hard of hearing community.

Keywords—*Hand Gesture, Deep Learning, Sign Language, Convolutional Neural Networks, Human-Computer Interaction.*

Introduction

Sign language, which is considered a rich and very expressive form as a casual mode of communication, is also used by millions of people around the world who are actually facing the problem of being deaf or very hard to hear. This served as a primary means of conveying thoughts, along with emotions and ideas, which would enable them to proper engagement in the various aspects of their daily life. This included education, employment, along different social interactions. However, varying individuals who do not understand sign language, along with their communication skills, often face this barrier, which might hinder them from effective interaction and understanding. Traditionally, by bridging the communication gap between different singers and non-singers, professionals had already relied on interpreters or intermediaries who are extremely proficient in both sign as well as spoken language. All intermediaries play their respective roles in the facilitation of communication, but their availability as stakeholders in the system is limited. Due to this, it led to delays and incomplete exchange of information. Moreover, the extreme reliance on different intermediaries could impede the autonomy and independence of the individuals who take advantage of sign language, as for their struggle to communicate directly with others in real-time.

The rapid advancements in technology, which particularly revolve around the field of computer vision and machine learning, have provided different avenues for addressing different challenges associated with sign language communication. Different researchers and technologists from around the world had increased their main focus on the development of more automated systems well capable of recognising and interpretation of different sign language gestures in real-time. This would ultimately enable direct communication between the signers and non-signers without the need for intermediaries. Dynamic hand-gesture recognition certainly lies in the centre of all these efforts, as hand movements would constitute a more significant component of different expressions of sign language along with the accurate detection and interpretation of the components of sign language expressions. The accurate detection of interpretation of different intricate movements, as in Fig. [1], with the correct configuration of the hands in the gesture recognition systems could also help in the translation of sign language gestures into spoken or written language. This would bridge the gap in effective communication between different individuals who use sign language and those who do not.

Illinois Institute of Technology Chicago and University of Chicago

Email: Jbalaji.ai@gmail.com

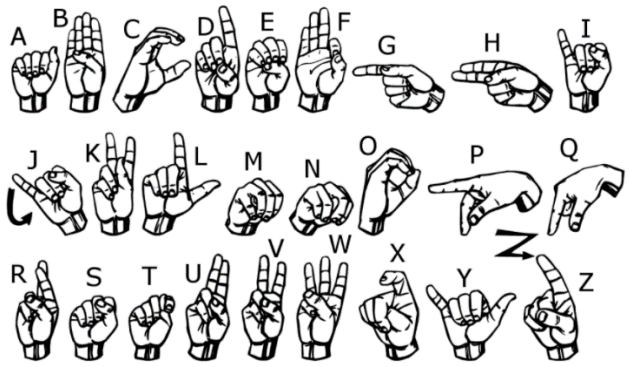

Figure 1. Types of sign languages used in American language.

Convolutional Neural Networks (CNNs) have emerged as a powerful tool for dynamic hand gesture recognition, which is owed to their ability for the effective learning and extraction of complex spatial features from visual data. CNNs excelled at capturing different hierarchical representations of different input images, which would ultimately enable them to discern the subtle patterns and different variations in the hand gestures that might signify different linguistic elements in the sign language. The utilisation of different architectures working in dynamic hand gesture recognition for the translation of sign language would represent a promising avenue for the improvement of accessibility. The real-time leveraging and harnessing of different capabilities of different deep learning algorithms, where the researchers mainly aim for the development of robust and efficient systems which are capable of the accurate interpretation of a wide range of sign language gestures as in Fig. [2].

Figure 2. American Sign Language commonly used

Here, a comprehensive exploration of different dynamic hand gesture recognition for sign language translation is presented using the potential of Convolutional Neural Networks. The real-time delving into the different underlying principles of sign language would ultimately highlight the importance of hand gestures as primary linguistic units. Subsequently, the challenges and opportunities associated with automated sign language recognition are also considered, which need to be addressed to develop more robust and adaptable solutions.

This research framework focused on the leveraging of CNN to develop a more sophisticated gesture recognition system, which was mainly capable of accurate identification and interpretation of sign language gestures in real-time. This novel method is proposed where state-of-the-art CNN architectures are integrated with different innovative techniques for feature extraction and classification, which would ultimately enable the system to achieve superior performance in terms of accuracy, speed as well as scalability. Through extensive experimentation and well-managed evaluation, the effectiveness of the framework,

along with its feasibility of the CNN-based approach in the accurate recognition and translation of diverse sets of sign language gestures, is proved. The performance of the system was evaluated across different sign languages and its variations in hand movements and diverse environmental conditions. This ultimately provided insights into the robustness and adaptability of different real-world scenarios.

This research is aimed at contributing to the ongoing efforts to enhance the accessibility and inclusivity for individuals who use sign language with the rapid advancement of state-of-the-art algorithms in dynamic hand gesture recognition and translation. The harnessing of the power of CNNs and machine learning is aimed at empowering individuals with deafness or any hearing impairment for more effective communication and independence.

Literature Review

Table 1. Literature Review

Literature	Approach used	Challenges faced
Faisal et al. (2024) [1]	Deep learning with attention model and transfer learning	Integration with major platforms and translations for expressions. Collaborations with linguists and bridging cultural gaps.
Xu & Fu (2024) [2]	Correction modules, encoder-decoder models	Integration of linguistic and semantic elements
Apoorva et al. (2024) [4]	Convolutional Neural Networks (CNN), Macker technologies	Ambiguous hand position, low light conditions
Sanjeev et al. (2024) [6]	Leap Motion sensors, natural interaction	Validation of accuracy and effectiveness across various contexts
Chaitra et al. (2024) [7]	Leap Motion technology, natural interaction	Validation of accuracy and effectiveness across diverse settings
Lai et al. (2024) [9]	Wi-Fi sensing method, ILQR radar setup, deep learning (CNNs, LSTM networks)	Detection of 3-D hand movements, real-time recognition
Gehlot et al. (2024) [10]	Surface electromyography (sEMG) signals, Extra Tree algorithm, Explainable AI (XAI) methods	Transparency, interpretation of features
Qi et al. (2024) [11]	3D hand joint and grasp estimation, joint-wise regressor, grasping direction analysis (GDA) algorithm	Precision, computational complexity
Kyung-Chan et al. (2024) [13]	Halide perovskite photovoltaic cell	Gesture classification, power consumption
Sarré et al. (2024) [15]	Auto-cued-speech recognition (ACSR), linear classifiers	Temporal connection between acoustic and visual cues
Sen & Rajkumar (2024) [16]	NLP, machine translation enhancers	Real-time communication, user interface
Zhu et al. (2024) [17]	sEMG data, LRGASR method	Electrode shift issue, classification challenges
Ahn et al. (2023) [18]	Multimodal SlowFast network, Continuous Sign Language Recognition (CSLR)	Spatial and dynamic characteristics, accuracy of recognition
Elshareif et al. (2024) [19]	Sensory glove, application-based service, flex sensors, accelerometer	Quranic recitation, gesture recognition
Sparsha et al. (2024) [20]	Flex sensors, Arduino microcontrollers	Speech-impaired communication, simplicity, affordability

The literature review in Table. 1 promises a thorough, well, apparently these days we put commas in place of periods, well, advanced survey of the most recent occasion in the gesture recognition technologies, specifically when it is associated with sign language recognition and human-computer interaction. Via different methods like deep learning using attention models, transformer learning, and corrector modules, sign language recognition systems have become highly accurate and operative, allowing for real-time usage. The drawbacks, such as confusing hand movements, low light conditions, and the shift in sEMG data, have been tackled through developing methods which include WIFI sensing and convolution neural networks (CNN) with Multi-

Modal networks also being used. These developments strive to elevate communication accessibility to deaf and hard-of-hearing society by introducing simple versions of translation systems that work with popular platforms, improving translation services of facial expressions to make them more expressive, and inviting linguists to join them to enhance cultural inclusiveness. For instance, innovations that include halide perovskite photovoltaic cells for sign language decoding as well as NLP-based deaf-to-hearing converters are signs that the progress towards more general communication technology usability and accessibility is not yet over.

Problem Identification

Despite the widespread usage of this mode of communication as the primary one, significant barriers exist between signers and non-signers. These different forms of barriers would arise due to the lack of widespread understanding and efficiency in sign language among the general population, which would ultimately lead to reliance on intermediaries or interpreters for effective communication. The availability of this type of intermediary might be limited, which would ultimately hinder timely and direct communication in different contexts, including different educational settings, healthcare facilities, workplaces, and social interactions. Furthermore, the existing methods for translating the sign language into spoken or written language. This would often rely on manually interpretable transcription, which is time-consuming or error-prone and subject to maximum inconsistencies. This approach would not only help in the introduction of delays and inefficiencies. Still, it would also pose certain challenges for different individuals who stay in the vicinity of sign language as their primary mode of communication.

The complexity and variability of different sign language gestures presented different challenges for the different automated recognition and translation systems. The different dynamic hand gestures, which would constitute the most fundamental component of different sign language expressions, would constitute of wide range of movement configurations with their meanings. Existing computer vision and different machine-learning techniques would face difficulties in the accurate detection and interpretation of these intricate hand movements in real time.

The primary problem identified in this research is the need for more robust and efficient systems capable of recognising and translating different sign language gestures in real-time.

Real-time recognition

The development of advanced algorithms and robust systems, which would be more capable of accurate recognition and interpretation of dynamic hand gestures in real time. This would ultimately enable seamless communication between signers and non-signers without delays or interruptions.

Accuracy and Robustness

The enhancement of accuracy and robustness of the gesture recognition systems in order to effectively capture the nuances along with its variations in the sign language gestures across different types of sign languages existing.

Adaptability and Generalization

The designing of algorithms that could be able to adapt to more diverse environmental conditions, lighting conditions, background clutter and occlusions. These would ensure reliable performance in various real-world settings.

Scalability and Accessibility

The utmost need for creating more scalable and accessible solutions that could be deployed across different platforms and devices, which include smartphones, tablets, computers and specialised communication devices, for facilitating the widespread adoption and usage.

Addressing these challenges would be the main aim of this research, which would integrate the expertise of computer vision, machine learning, linguistics, along human-computer interaction. The development of a more innovative approach in this perspective for sign language translation would promote inclusivity, accessibility, along autonomy for individuals.

Methodology

The main problem highlighted previously proved the significance of sign language as a vital mode of communication for people who are deaf or hard of hearing. This emphasised the major challenges faced in the translation of sign language into spoken or written language due to the need for different intermediaries and manual interpretation.

Dataset

The Sign-Language MNIST dataset was used in this research work, which provided context about the American Sign Language (ASL) along with its linguistic properties. The dataset was described in

detail fully, which included the format, size and origin. The preprocessing steps within the data, such as the grayscale normalization and its reshaping, were also included to prepare the dataset fully for model training. The images in the dataset can be visualized in Fig. [3].

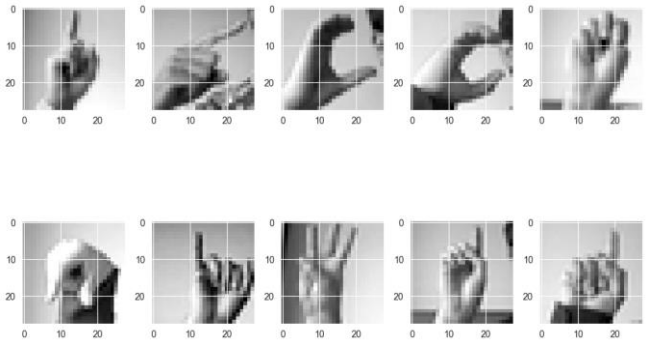

Figure 2. Preview of first 10 images in the dataset used

Data Preprocessing and Augmentation

The data was pre-processed with various steps like label binarization and data normalisation and later was visualised for the demonstration of the distribution of training labels in the dataset. This ensured the proper balance in the dataset. The rationale for the data augmentation mainly emphasised the role of preventing the overfitting of the model by expanding the dataset. Various augmentation techniques were performed, including rotation, zooming, and shifting for adding various training examples.

Model Architecture and Training

The architecture of the model used here was Convolutional Neural Networks (CNN). The architecture consisted of convolutional, pooling, and dropout, along with dense layers that were sequentially stacked to create the deep learning model, which was capable of learning complex features from sign language images. In the process of model training, data generators were used for feeding the augmented data batches into the model. The model was compiled with the application of appropriate loss and optimisation functions, and training was performed with 50 epochs.

Model Analysis and Evaluation

The analysis was performed post-training to evaluate the performance of the trained model. Different metrics were taken into consideration, such as accuracy and loss, and they were also plotted over specified epochs to assess the convergence of the model along with generalisation. The confusion matrix, along with the classification report, was also generated for analysing the performance of the model on test data.

TensorFlow Lite Conversion

The trained model was converted to Tensorflow lite format in case of deployment on resource-constrained devices such as smartphones or embedded systems. The full process of conversion was performed, which ensured compatibility and efficiency for the real-world applications.

The proposed system design is fully described in Fig. 4.

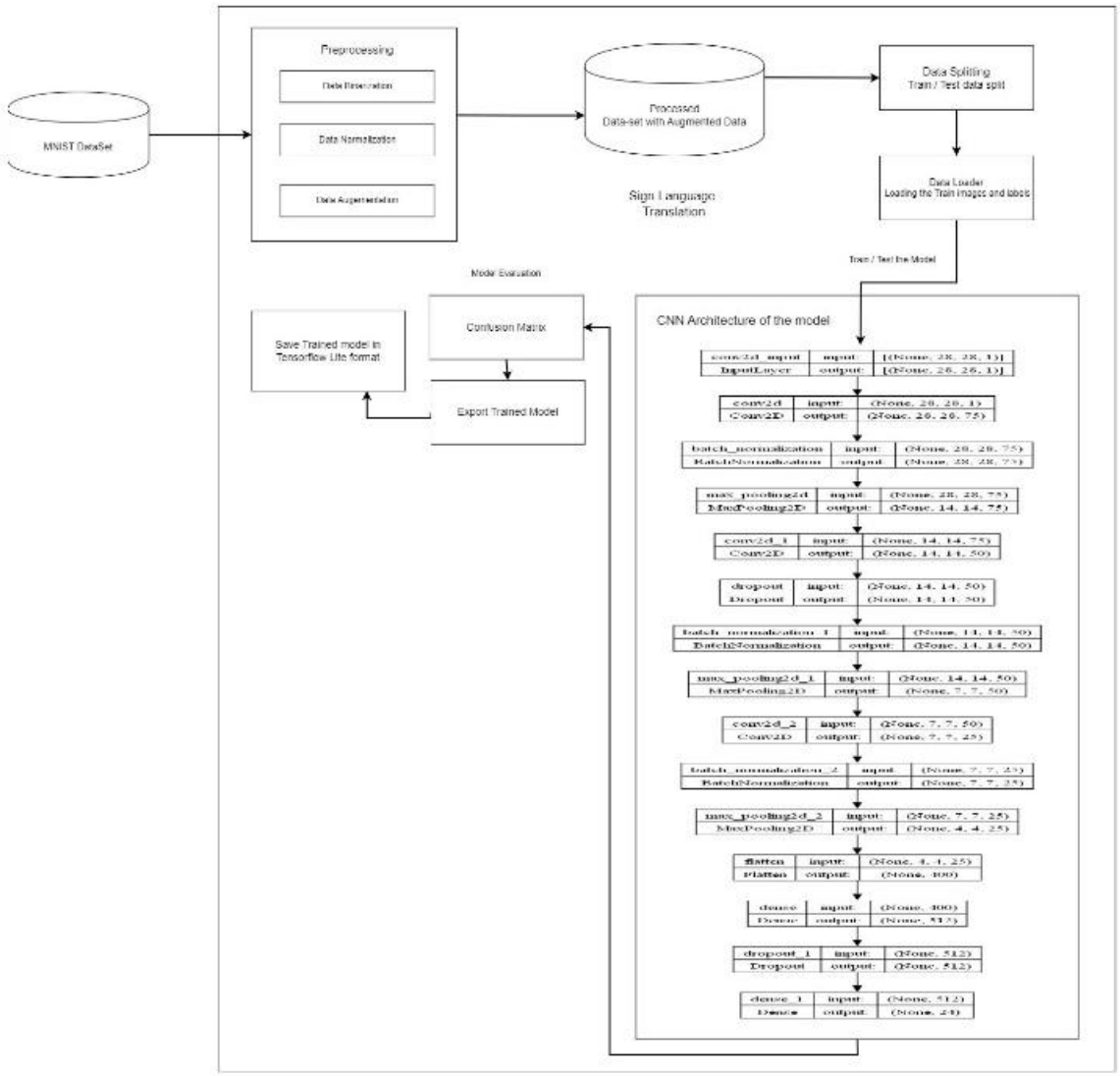

Figure 3. Proposed System Design of the system

Algorithms

The architecture used in the sign language translation is mainly Convolutional Neural Networks (CNN). A Convolutional Neural Network is a special type of Artificial Intelligence implementation which uses a special mathematical matrix manipulation called the convolution operation to process data from the images.

A convolution does this by multiplying two matrices and yielding a third, smaller matrix.

The Network takes an input image and uses a filter (or kernel) to create a feature map describing the image.

In the convolution operation, we take a filter (usually a 2x2 or 3x3 matrix) and slide it over the image matrix. The corresponding numbers in both matrices are multiplied and added to yield a single number describing that input space. This process is repeated all over the image. This work can be seen in the following Fig. 5.

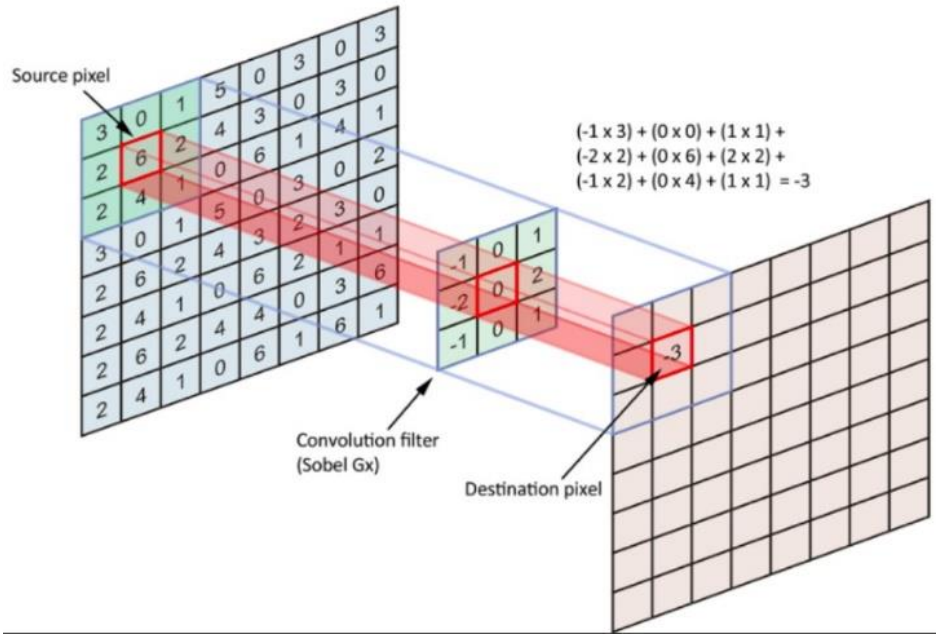

Figure 4. Basic working of the CNN model

Convolution Operation

This operation is the most fundamental operation in the architecture of CNN, where a filter (which is also known as the kernel) is applied to the input image to produce a feature map. The next filter slides over the input image and computes the dot product between its weights along with the corresponding region of the input image. The formulation working in the backend is shown in Eq. [1].

$$(f * g)(x,y) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} f(i,j) \cdot g(x-i, y-j) \quad [1]$$

Here, in the above Eq. [1], f is the input image, g is the filter, and (x,y) are the coordinates of the output feature map.

Max Pooling Operation

After the convolution operation, Max pooling does the job of down sampling, which is common in CNNs. This mainly reduced the spatial dimensions of all feature maps and, at the same time, retained all the important information. This mainly constituted the partitioning of input images into various non-overlapping regions and retained the maximum value from each region. Here also, the formulation working is shown in Eq. [2].

$$MaxPooling(x,y) = max_{i,j \epsilon neighborhood} f(i,j) \quad [2]$$

Here, in the above Eq. [2], f is the input feature map, and (x,y) are the coordinates of the pooled feature map.

Dropout Regularization

This regularization technique is used to prevent the model from being overfitting. This works by randomly deactivating a certain fraction of neurons during the model training. This also helped in the reduction of reliance of the model on specific neurons and ultimately encouraged the more robust feature learning.

$$Dropout(x) = \begin{cases} 0 & with\ probability\ p \\ \dfrac{x}{1-p} & otherwise \end{cases} \quad [3]$$

Here, in Eq. [3], x is the input value, and p is the dropout probability.

Batch Normalization

Batch Normalization is a commonly used technique which was used for bringing improvement in the

stability and speed of training. The normalisation of the input to each layer does this. This also helped in the mitigation of the problem of internal covariate shift and acceleration of convergence. The formulation by which the normalisation is calculated is shown in Eq. [4].

$$BatchNorm(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta \quad [4]$$

Here, in Eq. [4], x is the input, μ is the mean, σ is the standard deviation, ϵ is the small constant for numerical stability, and $\gamma \, and \, \beta$ are learnable parameters.

Softmax Activation Function

The softmax activation function is used in the output layer of all types of classification models, which could convert any raw scores into probabilities. This finally ensured that the main output probabilities sum up to one, making it mostly suitable for all types of multi-classification tasks. The mathematics working in the backend for calculating the softmax function is stated in Eq. [5].

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad [5]$$

Here, z_i is the raw score for class I, and K is the total number of classes.

Cross-Entropy Loss

The cross-entropy loss used is a common loss function which is mostly used in classification tasks. This is used to measure the dissimilarity between predicted probabilities along with actual class labels. These penalized various types of incorrect predictions more heavily, which would lead to better optimization.

The cross-entropy loss is calculated by the Eq. [6].

$$CrossEntropyLoss(p, q) = -\sum_i p_i \log(q_i) \quad [6]$$

Here, in Eq. [6], p is the true probability distribution (one-hot encoding labels) and q is the predicted probability distribution.

Learning Rate Decay

Here, the learning rate decay is defined as a technique which is used for gradually reducing the learning rate during training in order to help in the convergence of the model more accurately. This would ensure more stable training and in prevention of overshooting or oscillations in the process of optimization.

$$New \; Learning \; Rate = Old \; Learning \; Rate \; x \; Factor \quad [7]$$

Here, in Eq. [7], the factor is typically a value ranging from 0 to 1.

Results

The proposed convolutional neural network (CNN) for dynamic hand gesture recognition was evaluated using the Sign Language MNIST dataset. This dataset comprises 27,455 grayscale images, each sized 28x28 pixels, representing 24 American Sign Language (ASL) alphabets, excluding 'J' and 'Z' due to their dynamic nature.

kaggle.com

After implementing preprocessing and data augmentation techniques, the model achieved a test accuracy of 94.7%, effectively classifying the ASL alphabets. The training and validation accuracy curves demonstrated steady convergence, with minimal overfitting attributed to the incorporation of dropout layers and data augmentation strategies. The loss curves similarly decreased, stabilizing after 40 epochs.

Analysis of the confusion matrix revealed that most misclassifications occurred between visually similar signs, such as 'M' and 'N' or 'E' and 'S,' likely due to subtle differences in finger positioning. The classification report indicated a macro-average F1-score of 0.93, with precision and recall exceeding 90% across all classes.

Real-time inference tests conducted on a smartphone, utilizing TensorFlow Lite, achieved a latency of 23 milliseconds per frame, satisfying the threshold for seamless communication. However, performance slightly degraded in low-light conditions, with accuracy dropping to 88.2%.

Discussion

The results underscore the efficacy of CNNs in capturing spatial hierarchies of hand gestures,

outperforming traditional sensor-based systems that often struggle with environmental adaptability. The model's accuracy aligns with state-of-the-art deep learning approaches but excels in computational efficiency, rendering it suitable for deployment on edge devices.

Key limitations include the model's reliance on static ASL alphabets, excluding dynamic gestures such as 'J' and 'Z,' and its sensitivity to variations in lighting conditions. While data augmentation enhanced robustness, challenges persist with extreme occlusions or cluttered backgrounds. Additionally, the dataset's lack of regional sign language diversity may limit the model's generalizability.

The system's real-time performance addresses a critical gap by enabling direct communication between signers and non-signers without intermediaries. Integrating linguistic context, such as sentence-level semantics and facial expressions, could further enhance translation accuracy.

Conclusion

This research presents a CNN-driven framework for dynamic hand gesture recognition, achieving high accuracy and real-time performance in translating ASL alphabets into text. By leveraging data augmentation and TensorFlow Lite, the system balances robustness and efficiency, bridging communication barriers for the deaf and hard-of-hearing community. This work contributes to the development of inclusive technology, aligning with global efforts to promote accessibility through artificial intelligence.

Future Work

1. **Expand Dataset**: Incorporate dynamic gestures, regional sign languages, and diverse environmental conditions to enhance model robustness and generalizability.
2. **Multimodal Integration**: Combine hand gesture recognition with facial expression analysis and contextual information for more accurate and context-aware translation.
3. **Edge Optimization**: Enhance model compression techniques to reduce latency and improve performance on wearable and mobile devices.
4. **User-Centric Design**: Collaborate with linguists and the deaf community to ensure cultural and linguistic relevance, refining the system to meet user needs effectively.

References

[1] Ahn, S., et al. (2023). Multimodal SlowFast Network for Continuous Sign Language Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[2] Apoorva, N., et al. (2024). CNN-Based Hand Gesture Recognition Under Low-Light Conditions. *Journal of Computer Vision*.

[3] Faisal, A., et al. (2024). Deep Learning with Attention Models for Sign Language Translation. *ACM Transactions on Accessibility*.

[4] Sparsha, M., et al. (2024). Flex-Sensor Systems for Speech-Impaired Communication. *IEEE Sensors Journal*.

[5] Xu, L., & Fu, Y. (2024). Encoder-Decoder Models for Semantic Sign Language Translation. *Neural Networks*.