

Sign Language Recognition Algorithms Using Hybrid Techniques

Shakir Hussain Naushad Mohamed, Hao Feng Chan, Dexter Sing Fong Leong, Wui Chung Alton Chau
School of Engineering, Deakin University, Australia

Andi Prademon Yunus
Telkom University, Indonesia

Takao Ito
Hiroshima University, Japan

Zheng Cai, Xinjie Deng, Yit Hong Choo*
Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Australia
**Email: y.choo@deakin.edu.au*

Abstract

Sign language recognition is a vital tool for enabling communication with individuals who are hearing impaired. This paper proposes a custom gesture recognition framework designed specifically for sign language interpretation. The proposed model incorporates a deep learning approach trained on a custom dataset. The system achieves robust recognition of complex gestures while maintaining efficiency. This framework emphasizes adaptability to variations in sign language styles.

Keywords: Sign language recognition, pose estimation, gesture recognition, deep learning

1. Introduction

Sign language is an important communication tool for the hearing and speech-impaired community, yet only a small percentage of people know how to use it, creating significant challenges for effective communication. An Australian census revealed that only slightly more than 16,000 people use Auslan, the sign language of the country, highlighting the limited reach and understanding of sign language among the general population[1]. Individuals who rely on sign language often face difficulties when interacting with those who are not proficient in its use. The language encompasses hand gestures, facial expressions, and body movements, providing a visual mode of communication distinct from spoken language.

Effective communication is essential in every society, but for individuals with speech and hearing impairments, it remains a persistent barrier. This gap is especially evident in sectors like education, healthcare, and public services, where individuals from the hearing- and speech-impaired community struggle to communicate with others who do not understand sign language. Traditional communication methods, such as gestures, may be slow and limited in conveying complex ideas, adding to the frustration of those who rely on them.

While sensor-based systems have been explored as potential solutions[2], they are often expensive and complex to implement. In contrast, vision-based systems, which utilize video feeds to capture sign language gestures, have gained popularity due to their accessibility and scalability. These systems are more affordable and less intrusive, as they do not require specialized hardware such as sensor gloves. Using deep learning techniques to process and recognize hand gestures, vision-based systems are a more practical solution to the communication challenges faced by the hearing- and speech-impaired community.

Despite advancements in sign language recognition, challenges persist, particularly in accurately recognizing gestures that vary across individuals. Differences in hand shapes, orientations, and signing styles make it difficult to build robust recognition systems. A key challenge is designing models that generalize well across different signers while maintaining high accuracy. One study successfully demonstrated the application of CNNs in building a sign language recognition system for Indian Sign Language (ISL)[3], demonstrating its potential for accurate communication solutions.

This paper presents a vision-based sign language recognition model that uses a Hybrid CNN-LSTM architecture with an attention mechanism. The model combines the strengths of Convolutional Neural Networks (CNNs) for feature extraction, Long Short-

Term Memory (LSTM) networks for sequential learning, and an attention mechanism to focus on key image regions. It is trained on a custom dataset of American Sign Language (ASL) gestures to achieve accurate recognition of signs. The aim is to enhance communication between the hearing- and speech-impaired community and the public.

To evaluate the model's effectiveness, it is benchmarked against SqueezeNet[4], a lightweight CNN architecture renowned for its efficiency in image classification tasks. Specifically, this study focuses on detecting the ASL hand signs for the numbers “1,” “2,” and “3.” The hybrid model’s performance and efficiency are compared to that of SqueezeNet, providing insights into the benefits of combining CNN, LSTM, and attention mechanisms for sign language recognition.

2. Methodology

The custom dataset contains images of static ASL gestures. These images are preprocessed through resizing, normalization, and augmentation techniques such as rotation and color jittering to ensure the model generalizes well across different conditions. The dataset is split into training, validation, and testing sets, and the performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrices.

For training, the CNN component extracts the relevant features, and the LSTM unit captures any sequential dependencies in the features. The attention mechanism ensures the model focuses on the most informative parts of the input images. The output of the model is classified into predefined classes based on the extracted features, with the training process optimized using stochastic gradient descent and cross-entropy loss for multi-class classification.

The performance of the model is compared with that of benchmark models, specifically SqueezeNet, to evaluate its effectiveness in detecting static gestures. Performance metrics, including mAP, precision, recall, F1-score, and accuracy, are computed to assess the accuracy and efficiency of the model in gesture classification.

2.1. Dataset preparation

For this study, a custom dataset was created using Roboflow[5], focusing on images representing static sign language gestures. Roboflow was instrumental in data annotation, augmentation, and preprocessing.

The dataset consists of images representing different 3 sign language classes of the numbers 1, 2 and 3 as shown in Fig.1, Fig.2 and Fig.3, respectively and was split into three subsets:

- **Training Set:** Comprising 69% of the total data (1,164 images), this subset was utilized for model training.
- **Validation Set:** Comprising 19% of the total data (324 images), this subset was reserved for hyperparameter tuning and monitoring model performance during training.
- **Test Set:** Comprising 11% of the total data (192 images), this subset was strictly used for the final evaluation of the trained models.

During preprocessing, images were auto oriented to ensure consistent alignment and resized to a uniform dimension of 640x640 pixels, aligning with the input requirements of the benchmarking models used in the later phases of the study. The dataset was exported in COCO format, a widely accepted standard for object detection tasks, facilitating compatibility with various deep learning frameworks.



Fig. 1. Hand signal of number 1 in ASL



Fig. 2. Hand signal of number 2 in ASL



Fig. 3. Hand signal of number 3 in ASL

2.2. SqueezeNet

SqueezeNet is a compact convolutional neural network specifically designed to balance high accuracy with minimal model size. It employs "fire modules," which involve an initial squeeze layer utilizing 1x1 convolutions to reduce the input channels, followed by an expand layer with a combination of 1x1 and 3x3 convolutions for enhanced feature extraction. This architectural approach significantly reduces the parameter count and computational demands while maintaining robust performance. Therefore, SqueezeNet serves as an optimal benchmark model for comparative analysis.

2.3. Base Convolutional Neural Network (CNN)

A custom CNN architecture was developed to serve as the backbone for feature extraction and classification, specifically designed to optimize feature capture, depth, and regularization for handling the hand sign dataset. The architecture is composed of two primary components: feature extraction layers and fully connected layers dedicated to classification. The feature extraction module comprises four convolutional blocks with increasing depth, ranging from 64 to 512 channels. Early layers employ 5x5 kernels to capture spatial details effectively. Each convolutional block incorporates batch normalization and ReLU activation functions for stability and non-linearity. Max pooling is applied to reduce spatial dimensions, which aids in focusing on prominent features. Flattened features from the convolutional blocks are passed through the fully connected layers, beginning with an input size of 512x4x4. Regularization is implemented using a dropout rate of 0.5 to mitigate overfitting. The final layer produces probabilities for the three hand signal classes.

2.4. CNN with LSTM and Attention mechanism

The hybrid CNN architecture combines the base CNN mentioned in the previous section with an LSTM module, and an attention mechanism

CNNs extract spatial features from images, but these features may still have inherent relationships. LSTMs help model these dependencies, improving feature interpretation[6].

The attention mechanism focuses on the most relevant features by assigning weights to the outputs of the LSTM. By computing a weighted sum of the LSTM outputs, the attention mechanism ensures that the model prioritizes key information while suppressing less important features[7].

Finally, fully connected layers are used to perform the final classification, with dropout applied for regularization to reduce overfitting. This model is designed to efficiently handle static gesture recognition tasks.

2.5. Training and validation

Training: The training process begins by setting up a Stochastic Gradient Descent (SGD) optimizer and a cyclic learning rate scheduler to adjust the learning rate during training. The models are trained over 25 epochs, where the cross-entropy loss is computed, and the weights are updated through backpropagation. The training loop includes tracking of training accuracy, loss, and mean average precision (mAP) for each epoch. After every epoch, the model is evaluated on the validation set to check if the validation loss improves, saving the best model based on the lowest validation loss.

Testing: After training, the performance of the model is evaluated on the test set and various metrics were computed, including the confusion matrix, F1-score, ROC curve and overall accuracy.

2.6. Performance Metrics:

The effectiveness of the models were evaluated using standard metrics, mathematically defined as follows[8]:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (2)$$

$$\text{F1 - Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3)$$

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (4)$$

As shown in Eq. (4), accuracy is the overall proportion of correctly classified instances. As in Eq. (1), precision indicates how much of the result are actual positives out of the predicted positives. As indicated in Eq. (2), recall calculates how much of the actual positives are labelled as true positives. F1-Score as mentioned in Eq. (3), is a function of precision and recall and will provide a value to balance between precision and recall.

These metrics provide a robust assessment of the performance of the models, performance, balancing its ability to correctly classify signs while minimizing false positives and false negatives.

The input data undergoes a series of transformations to enhance model robustness and generalization. These transformations include resizing the images to a consistent size, random horizontal flipping, random rotation, and random colour jittering, which adjusts brightness, contrast, saturation, and hue. These augmentations are applied to all the models, ensuring that the models can effectively handle variations in the input

data, such as different orientations and lighting conditions, ultimately improving their performance and ability to generalize to unseen data.

3. Results and Discussion

This section presents a comparative analysis of the base CNN, the proposed hybrid CNN-LSTM with attention model, and SqueezeNet for ASL sign recognition of the numbers “1”, “2” and “3”. All experiments were conducted on a system equipped with an NVIDIA GeForce RTX 2070 GPU, an Intel Core i7-10750H CPU operating at 2.60 GHz, and 16 GB of RAM.

The performance of the three models is summarized in Table 1, using accuracy, precision, recall, and F1-score as evaluation metrics. Additionally, the relationship between training loss and epochs was used as the primary evaluation metric, providing insight into the convergence behavior of each model during training.

Table 1. Performance metrics of CNN, Hybrid CNN and SqueezeNet

Metric	SqueezeNet	Base CNN	Hybrid CNN
Accuracy	89%	89%	95%
Precision	90%	88.66%	95%
Recall	88.33%	88.33%	95%
F1-Score	88.33%	88%	95%
Final mAP	96.93%	96.25%	99.6%

From the results in Table 1, it is clear that the hybrid CNN-LSTM with attention model significantly outperforms the base CNN and SqueezeNet in all metrics. Its ability to integrate convolutional feature extraction with temporal pattern recognition, combined with attention mechanisms, allows it to focus on the most critical spatial regions, enhancing its precision and recall.

In contrast, the base CNN and SqueezeNet models perform similarly in terms of accuracy, but SqueezeNet exhibits slightly higher precision due to its efficient architecture, which minimizes redundancy in feature extraction. However, its recall and F1-score remain comparable to the base CNN, indicating a potential trade-off between efficient processing and model sensitivity to certain classes.

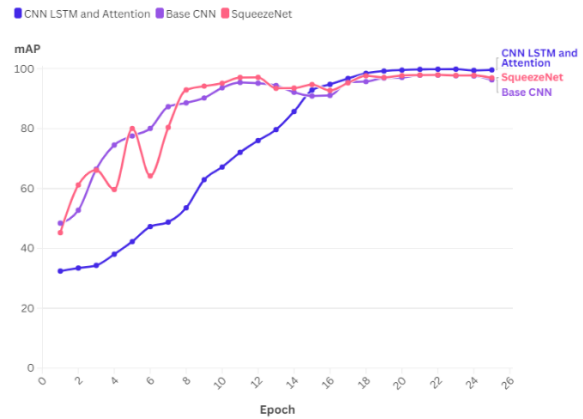


Fig. 4. Epochs vs mAP%

As shown in Fig.4, the mean Average Precision (mAP) scores across epochs of the SqueezeNet, Base CNN and the CNN with LSTM and attention mechanism models, further show the superiority of the hybrid CNN-LSTM with attention model. It consistently outperforms the base CNN and SqueezeNet, particularly in the later stages of training, achieving an mAP of 99.6% by the 25th epoch.

The ability of the hybrid model to achieve high mAP scores at later epochs highlights its strength in balancing precision and recall across varying confidence thresholds. By leveraging attention mechanisms, the model retains essential spatial details critical for distinguishing fine-grained features in the gestures.

In contrast, the base CNN achieves a respectable mAP but begins to level after 15 epochs, reflecting its limited capacity to generalize beyond certain levels of precision and recall. SqueezeNet achieves strong performance initially, peaking at epoch 10, but its mAP scores gradually decline thereafter, indicating a lack of robustness in maintaining balanced precision and recall across varying thresholds.

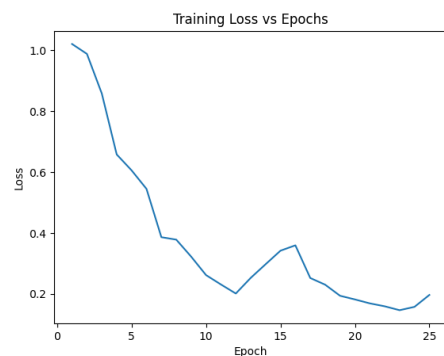


Fig. 5. Training loss vs Epochs Base CNN

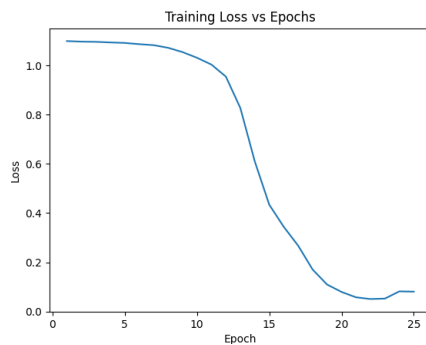


Fig. 6. Training loss vs Epochs Hybrid CNN

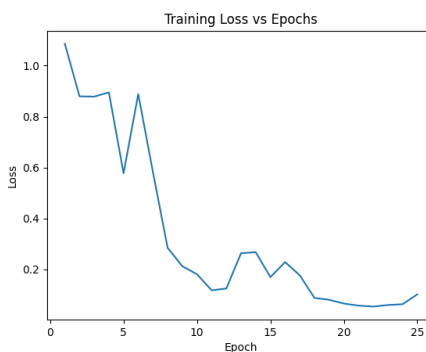


Fig. 7. Training loss vs Epochs SqueezeNet

Figures Fig. 5, Fig. 6 and Fig. 7 show that only the hybrid CNN model has a consistent steady decrease in loss, this can be attributed to its strong optimization capabilities and the combination of convolutional feature extraction, feature map sequence processing by the LSTM and attention mechanisms.

4. Conclusion

In this paper a hybrid deep learning model combining CNNs, LSTMs, and attention mechanisms for recognizing static hand gestures in American Sign Language (ASL) is proposed. The hybrid CNN model demonstrated strong performance across metrics such as accuracy and F1-score when compared to SqueezeNet. Data augmentation improved generalization, enabling adaptability to input variations.

However, limitations include the small size of the dataset and the lack of sequential data, which restricted the ability of the model to capture temporal dependencies and limited its robustness to complex or diverse gestures. These constraints highlight the need for larger, more comprehensive datasets and improvements in handling sequential data for better generalization and performance. Future work could focus on expanding the dataset, incorporating dynamic gesture recognition. Exploring multimodal approaches that integrate gesture recognition with other sensory data could further advance sign language detection systems.

References

1. D. F. Australia. "The Deaf Census: find out about Auslan users." <https://www.deafnessforum.org.au/the-deaf-census-find-out-about-auslan-users/> (accessed 2025/01/21, 2025).
2. A. M. Buttar, U. Ahmad, A. H. Gumaei, A. Assiri, M. A. Akbar, and B. F. Alkhamees, "Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs," *Mathematics*, vol. 11, no. 17, p. 3729, 2023. [Online]. Available: <https://www.mdpi.com/2227-7390/11/17/3729>.
3. A. R. Satish Kumar Alaria, Vivek Sharma, Vijay Kumar, "Simulation and Analysis of Hand Gesture Recognition for Indian Sign Language using CNN," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 10, no. 4, pp. 10-14, 2022, doi: <https://doi.org/10.17762/ijritcc.v10i4.5556>.
4. S. H. Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, Kurt Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size," presented at the ICLR, 2016. [Online]. Available: <https://arxiv.org/pdf/1602.07360>.
5. PoseEstimation. "Hand Signals Dataset." Roboflow. <https://universe.roboflow.com/poseestimation-nzxk0/hand-signals-m7ruz> (accessed 19/01/2025, 2025).
6. Y. He, C. Wen, and W. Xu, "Residual Life Prediction of SA-CNN-BILSTM Aero-Engine Based on a Multichannel Hybrid Network," *Applied Sciences*, vol. 15, no. 2, p. 966, 2025. [Online]. Available: <https://www.mdpi.com/2076-3417/15/2/966>.
7. Z. L. Jingyi Wang, Qiang Liu, Shu Wu, "Towards Accurate and Interpretable Sequential Prediction: A CNN & Attention-Based Feature Extractor," presented at the Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19), Beijing, China, 2019. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3357384.3357887>.
8. T. Abdullah All, E. M. Mahir, S. Akhter, and M. R. Huq, "Detecting Fake News using Machine Learning and Deep Learning Algorithms," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 28-30 June 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843612. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8843612>

Authors Introduction

Mr. Shakir Hussain Naushad Mohamed



He is currently an undergraduate student majoring in Bachelor of Electrical and Electronics Engineering (Honours) at Deakin University, Australia.

Mr. Hao Feng Chan



He is currently an undergraduate student majoring in Bachelor of Mechatronics Engineering (Honours) at Deakin University, Australia.

Mr. Dexter Sing Fong Leong



He is currently an undergraduate pursuing Bachelors Of Mechatronic Engineering (Honours) in Deakin University, Australia.

Mr. Chau Wui Chung Alton



He is currently an undergraduate student in Bachelor of Mechanical Engineering (Honours) at Deakin University, Australia.

Andi Prademon Yunus, Ph.D



He is an Assistant Professor at Telkom University and he received his PhD in Engineering from Mie University, Japan. His research focuses on applied and fundamental machine learning for motion and behavior computing. He also collaborates with industry partners to develop AI-based tools for language modeling and image analytics.

Dr. Takao Ito



He is Professor of Management of Technology (MOT) in Graduate School of Advanced Science and Engineering at Hiroshima University. His current research interests include automata theory, artificial intelligence, systems control, quantitative analysis of interfirm relationships using graph theory, and engineering approach of organizational structures using complex systems theory.

Mr. Zheng Cai



He is currently a PhD candidate at Deakin University's Institute for Intelligent Systems Research and Innovation (IISRI). His research interests include multiobjective optimisation algorithms such as metaheuristic algorithms and evolutionary algorithms for scheduling problems. He is also exploring the integration of machine learning with optimisation algorithms.

Ms Deng Xinjie



She is pursuing a PhD in Information Technology at the Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, with her research centered on creating lightweight deep learning algorithms for computer vision applications.

Dr. Yit Hong Choo



He has completed his PhD and is now a Research Fellow in Operations Analytics at Deakin University's Institute for Intelligent Systems Research and Innovation (IISRI), supported by the Rail Manufacturing Cooperative Research Centre (RMCRC). His research focuses on advanced multi-objective optimisation algorithms for complex maintenance scheduling in rolling stock. He collaborates with transportation industry partners to develop AI-based tools for video and image analytics.