Exploring MediaPipe Optimization Strategies for Real-Time Sign Language Recognition

Nguyen Phuoc Thanh[1], Nguyen Thanh Hoang[2], Tran Trong Tin[4], Le Ngoc Khanh[2], Huynh Hieu Nhan[4], Phan Huynh Thanh Binh[2], Vu Hoang Son Hai[3], and Hoang Ngoc Xuan Nguyen[2]

FPT University, Ho Chi Minh City, Vietnam
`thanhnpse171408@fpt.edu.vn`

Abstract. This research introduces an innovative approach to optimizing real-time sign language recognition (SLR) models while maintaining a relatively high level of accuracy. We propose a model that leverages the Mediapipe framework [14] for pose estimation and incorporates a simple Long Short-Term Memory (LSTM) network [21] for vocabulary prediction based on the actions performed by an individual captured through a camera.

By harnessing the capabilities of Mediapipe [14], we extract essential features from video frames, including bones, joints, bone motions, and joint motions. These features are then fed into the LSTM network [21], effectively capturing temporal dependencies and contextual information. To evaluate the performance of our system, we conducted experiments using two publicly available datasets [17][5], including the Green Screen RGB clips dataset from How2sign [5].

Our proposed method achieved state-of-the-art results on these datasets, demonstrating its effectiveness in accurately identifying sign language actions. Notably, our system exhibits scalability and simplicity of implementation when applied to real-time datasets, thanks to the efficient processing capabilities of Mediapipe [14]. Furthermore, we showcase the practical application of our system by presenting a live SLR demo that operates in real-time on Jupyter Notebook.

By combining the strengths of Mediapipe [14] and LSTM [21], our approach offers a promising solution for optimizing real-time sign language recognition models. Its performance, simplicity, and applicability make it a valuable contribution to the field, with potential applications in various domains such as assistive technology, communication systems, and accessibility tools.

Keywords: SLR · LSTM · ISL · MediaPipe · Multi-Modal Multi-Stream · Skeleton · Indian Sign Language · How2Sign.

1 Introduction

Sign language recognition (SLR) is a challenging task that has garnered significant attention in recent years, requiring the ability to accurately interpret deceptive hand motions, subtle body movements, and nuanced facial expressions

[11, 18]. This paper presents a novel approach to SLR that focuses on real-time implementation and optimization, aiming to enhance the precision and efficiency of sign language recognition models [4][8].

The proposed method leverages whole-body key points and skeleton-based techniques to address the inherent difficulties of SLR [18]. The dynamics of sign language are effectively captured by utilizing the Sign Language Graph Convolution Network (SL-GCN) [3] and the Separable Spatial-Temporal Convolution Network (SSTCN) [19]. These models incorporate advancements in whole-body pose estimation, enabling the accurate representation and analysis of the complex gestures involved in sign language communication. Furthermore, the study introduces the Skeleton Aware Multi-modal SLR framework (SAM-SLR) [18], which combines the strength of the skeleton-based approach with various modalities, including RGB and RGB-D data, to improve identification performance further.

The contributions of this study can be summarized as follows:

1. Introduction of a novel skeleton graph for SLR, utilizing whole-body key points and graph reduction techniques without additional annotation work.
2. Proposal of SL-GCN, an effective utilization of whole-body skeleton graphs to address the challenges of SLR.
3. Introduction of SSTCN, a state-of-the-art technique that accurately represents whole-body critical points more accurately than traditional 3D convolutions.
4. Demonstration of the SAM-SLR system for RGB and RGB-D-based SLR, incorporating data from multiple modalities and achieving top performance on the AUTSL dataset and the CVPR-21 Challenge on Isolated SLR.

To provide readers with a visual understanding of the proposed model, Figure 1 illustrates the concept of our Skeleton Aware Multi-modal Sign Language Recognition model. This model captures local and global motion information, extracting and utilizing it for accurate prediction 3.3. By integrating multi-modal data [5] and optimizing the architecture for real-time performance, our approach aims to significantly enhance sign language recognition capabilities.

The rest of this paper is organized as follows. Section 2 provides a comprehensive review of related work in sign language recognition, highlighting the limitations of existing approaches and motivating the need for our proposed method. Section 3 presents the details of our methodology, focusing on the model design and optimization techniques for real-time implementation. Section 4 describes the experimental setup, including dataset preparation, training procedures, and comparative evaluations 4.3 with other state-of-the-art models [8]. Section 5 discusses the results and provides insights into the strengths and limitations of our proposed approach. Finally, Section 6 concludes the paper by summarizing the contributions, discussing future research directions, and emphasizing our work's impact on sign language recognition.

Overall, this paper aims to advance the field of sign language recognition by utilizing whole-body key points, skeleton-based approaches, and multi-modal data to achieve real-time and optimized performance. We aim to develop more accurate and efficient sign language recognition systems through our innovative model and experimental evaluations.

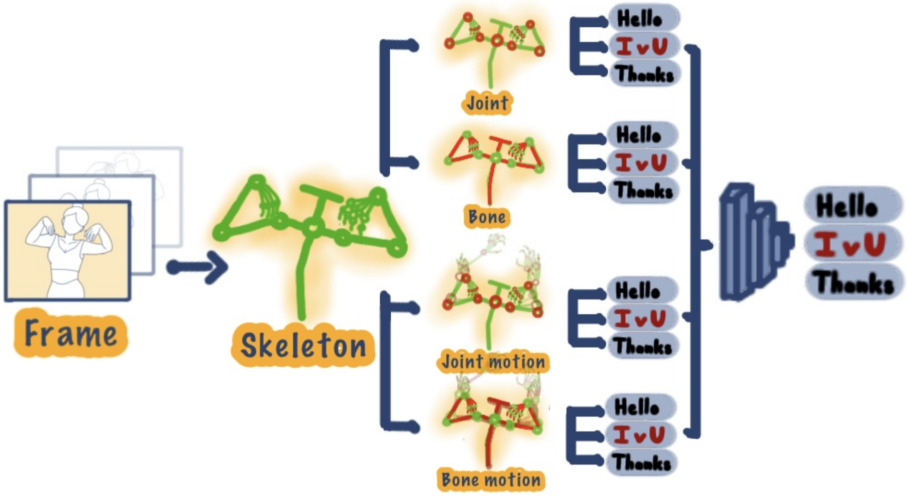

Fig. 1. Illustration of the proposed multi-modal for sign language recognition. The model leverages skeleton posing and incorporates diverse modalities to enhance recognition performance.

2 Related work

In recent years, significant research in hand sign detection has focused on utilizing skeleton-based approaches. This section overviews some relevant studies and applications in this area.

Sign language detection has seen significant recent research focusing on achieving real-time performance using skeleton data.

One notable study by Hu et al. [9] proposed a correlation network model to recognize continuous sign language gestures. The model effectively captured the temporal dynamics of sign language, allowing for accurate recognition of consecutive gestures. However, the complexity of the correlation network architecture could require significant computational resources.

Another significant work by Zhang et al. [12] focused on real-time hand gesture detection and classification using convolutional neural networks (CNNs). By leveraging CNNs, the researchers could detect and classify hand gestures in real-time, benefiting from the ability of CNNs to capture spatial information from skeleton data. This approach offered robust performance in detecting and classifying hand gestures. However, one potential drawback of CNN models is their limited ability to handle complex temporal dependencies in continuous sign language sequences.

Jiang et al. [10] presented a skeleton-based sign language detection study published in the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. Their approach involved extracting relevant features from skeleton data and employing deep learning techniques for classification. The proposed method enhanced the recognition performance by utilizing the structural information provided by skeleton data. However, a potential limitation is that accurate and high-quality skeleton data capture might be challenging to obtain in real-world scenarios.

In large-scale sign language recognition, Wang et al. [13] introduced a novel dataset and compared different word-level sign language recognition methods from videos. Although this study primarily focused on video-based approaches, it provided valuable insights into the challenges and potential solutions in sign language detection. Video-based methods offer the advantage of capturing spatial and temporal information, leading to a more comprehensive understanding of sign language. However, the computational intensity of video processing in real-time can limit practical implementation.

Chen et al. [2] proposed a Two-Stream Network for Sign Language Recognition and Translation. This approach combined the modalities of skeleton and video to improve the performance of sign language recognition and translation tasks. By leveraging both modalities, the model achieved enhanced accuracy and robustness. However, incorporating multiple modalities might increase the complexity and computational requirements of the system.

In summary, the existing research on sign language detection has explored various approaches, including correlation networks [9], CNNs, skeleton-based methods, and video-based methods[17]. Each system has advantages and limitations in capturing temporal dynamics, spatial information, structural cues, and scalability. Given the emphasis on real-time performance, this research will focus on utilizing skeleton data for real-time hand sign detection, taking into consideration the performance trade-offs of the existing solutions.

3 Methodology

3.1 Baseline

Sign language recognition is crucial for communication for individuals with hearing and speech impairments. It not only helps those with disabilities to communicate effectively but also has the potential to aid in the brain development of individuals without impairments [6]. Developing a model that accurately interprets sign language in real-time while maintaining high accuracy poses a significant challenge [4]. To address this, we propose a novel approach that utilizes skeleton pose estimation for real-time sign language recognition, as illustrated in Figure 2.

The proposed model employs sophisticated computer vision techniques to accurately estimate the skeleton poses based on the input video stream. It operates in real-time and effectively detects and tracks significant joints of the human body, including hands, elbows, and shoulders [11]. By extracting the precise spatial coordinates of these joints, a comprehensive skeleton representation is obtained, encapsulating crucial information for sign language recognition.

The skeleton-based representation allows us to focus on sign language gestures' spatial and temporal characteristics, enabling more accurate recognition. Instead of processing the entire video frame, we only need to analyze the skeleton coordinates, significantly reducing computational complexity and improving efficiency.

Machine learning algorithms, such as deep neural networks, are employed to recognize sign language. The model learns to establish associations between specific patterns and movements and their corresponding signs by training on a diverse dataset [17][5] containing various sign language gestures. Rather than focusing solely on gestures, the model utilizes skeleton pose estimation techniques to capture the underlying skeletal structure. During real-time recognition, the supplied skeleton coordinates serve as input for the model, predicting the associated sign language gesture.

The advantage of our proposed approach lies in its ability to perform recognition in real-time without compromising accuracy. By relying on the skeleton representation, our model eliminates the need for extensive image processing and allows for faster computation. Moreover, the skeleton-based approach ensures robustness to variations in lighting conditions, background clutter, and occlusions, thereby enhancing the overall correctness of the recognition results.

In conclusion, our sign language recognition model based on real-time skeleton pose estimation offers an efficient and accurate solution for interpreting sign language gestures. By leveraging the advantages of skeleton-based representation and employing machine learning algorithms, we enable real-time recognition while maintaining high accuracy and robustness. This model holds great potential for various applications, including assistive technologies, educational tools, and human-computer interaction systems for individuals with hearing impairments.

3.2 Skeleton Posing Model

To achieve real-time and precise skeleton posing, we leverage the capabilities of MediaPipe [14], an open-source tool renowned for its exceptional speed and accuracy. The MediaPipe [14] framework provides a comprehensive solution for computer vision and machine learning tasks, including robust and precise pose estimation.

By utilizing MediaPipe [14] pre-trained pose estimation model, we can efficiently detect and track critical body joints with high accuracy. This model leverages deep learning techniques to perform better in real-time skeleton recognition.

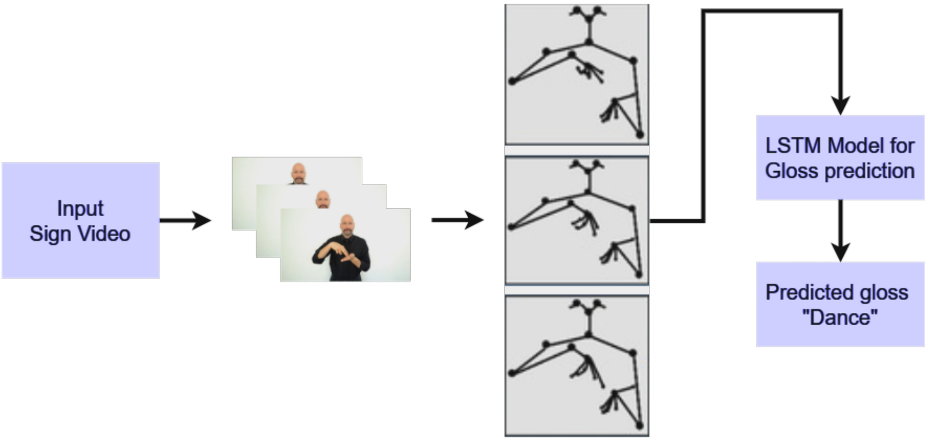

Fig. 2. Proposed Skeleton Pose for Real-Time Sign Language Recognition

The optimized implementation and architecture of MediaPipe [14] contribute to its remarkable speed and precision.

We rely on MediaPipe [14] pose estimation capabilities to obtain precise, real-time skeleton information for our sign language recognition model. The MediaPipe [14] model accurately detects and tracks the critical body joints, enabling us to estimate skeleton poses efficiently from the input video stream.

To capture the crucial details of sign language gestures, we specifically selected 34 points on the stance, 468 points on the face, 21 points on the left hand, and 21 points on the right hand. These fixed points correspond to significant landmarks that capture the intricate movements and positions of the signer's body.

Considering this comprehensive set of skeleton points, our model can capture the gross and fine-grained movements involved in sign language gestures. This complete representation enhances the accuracy of recognizing the signer's intent and overall recognition performance.

The combination of MediaPipe [14] pose estimation model, known for its precision and speed, along with the specific selection of skeleton points, forms a solid foundation for our sign language recognition model. This approach enables real-time recognition, ensuring the model accurately captures the subtle nuances of sign language gestures.

In conclusion, by leveraging MediaPipe [14] pose estimation capabilities and selecting specific skeleton points, our proposed sign language recognition model achieves real-time and accurate recognition of sign language gestures. This approach holds great promise for various applications facilitating communication for individuals with hearing impairments.

3.3 Multi-Modal Multi-Stream Approach

Based on the concepts 1 mentioned, we propose a multi-stream approach to address the challenges of recognizing sign language gestures based solely on skeleton coordinates 30 frames. We have identified that continuous motion-based signs often have similar skeleton coordinates over time, which can lead to ambiguity. We suggest implementing a multi-model ensemble approach to optimize this issue to provide real-time recognition results.

The first model initial idea of recognizing sign language gesture involves using the skeleton joint coordinates $j(x_i^t, y_i^t, z_i^t)$. Here, x_i and y_i represent the joint coordinates in a single image frame, while z_i denotes the confidence level associated with that particular joint. The index t indicates the image sequence's frame number or temporal aspect.

The second model introduces the concept of bone motion. We create vectors $b(x_j^t - x_i^t, y_j^t - y_i^t, z_i^t)$ based on the changes in bone positions as an action pose. These bone vectors aim to capture the movement patterns associated with different sign language words.

The third model utilizes a heatmap representation of the skeleton joint coordinates that changes over time across different frames. This approach helps distinguish between similar skeleton configurations from different actions. Additionally, we incorporate joint motion features by calculating the differences between the current frame and the previous frame, specifically by subtracting the coordinates of the current bone positions from the coordinates of the corresponding bones in the previous frame. We denote this joint motion as $jm(x_i^{t+1} - x_i^t, y_i^{t+1} - y_i^t, z_i^t)$.

The fourth model focuses on the differences between the current frame and the previous frame by calculating the coordinates of the bone vectors in the current frame based on the previous frame's bone vectors. Specifically, we calculate the bone motion as the differences between the bone positions in the current frame and the corresponding positions in the previous frame. We represent this bone motion as $bm(x_i^{t+1} - x_i^t, y_i^{t+1} - y_i^t, z_i^t)$.

We train these four models independently to recognize individual actions using the initial landmark points. Joint and bone motion calculations involve processing frames over time; thus, we propose using the initial model architecture while performing calculations on a sequence of 30 consecutive frames based on the series of experimental results 4.2.

Finally, we combine the predictions from the four models using an ensemble approach to generate the final recognition result. The prediction flow formula is as follows:

$$q = J \cdot \alpha_1 + B \cdot \alpha_2 + JM \cdot \alpha_3 + BM \cdot \alpha_4$$

Where J, B, JM, and BM represent the predictions from each respective model with different inputs, namely Joint, Bone, Joint Motion, and Bone Motion, and α_1, α_2, α_3, and α_4 represent the individual model weights.

By employing a multi-stream ensemble approach, we aim to enhance the accuracy and robustness of our sign language recognition system by leveraging different aspects of the skeleton data, including spatial information, bone motion, joint motion, and temporal dynamics. This approach could improve real-time recognition performance for sign language gestures.

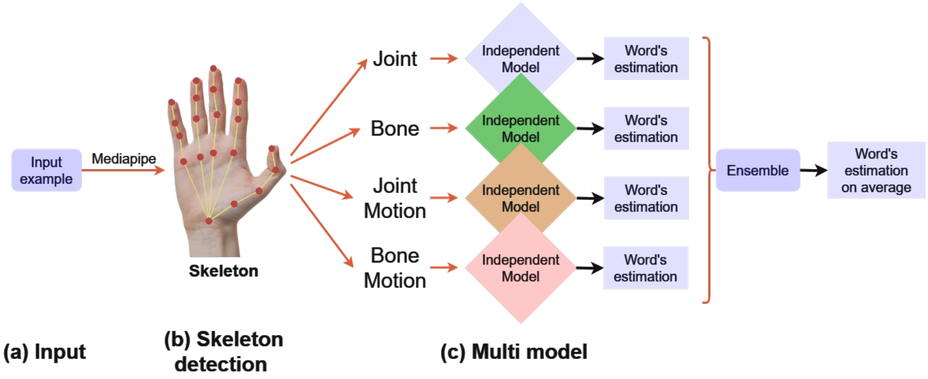

Fig. 3. Illustration of the model pipeline: Process visualization

3.4 Network Approach

LSTM: Long Short-Term Memory [21] is a Recurrent Neural Networks [16] type designed to address the vanishing gradient problem in traditional Recurrent Neural Networks [16]. A typical Long Short-Term Memory [21] unit consists of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell.

Dense and Dropout: The Dense layer, the fully-connected layer, creates abstract representations of input data by connecting neurons to every neuron in the preceding layer. The Dropout layer [20] is a regularization technique to enhance generalization and avoid excessive reliance on specific patterns. During training, the Dropout layer randomly deactivates input units with a certain probability, promoting robustness and preventing the model from overly fitting to the training data.

Activation Functions: The Rectified Linear Unit [1] activation function is a piece-wise linear function that will output the input directly if it is positive. Otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often performs better. The soft-max activation function [15] transforms the raw outputs of the neural network into a vector of probabilities, essentially a probability distribution over the input classes. It is often used as the last activation function of a neural network to normalize the output of a network to a probability distribution over predicted output classes.

The soft-max activation function [15] is defined as:

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

x_i is the input value of the i-th element in the input vector, and N is the number of elements in the input vector.

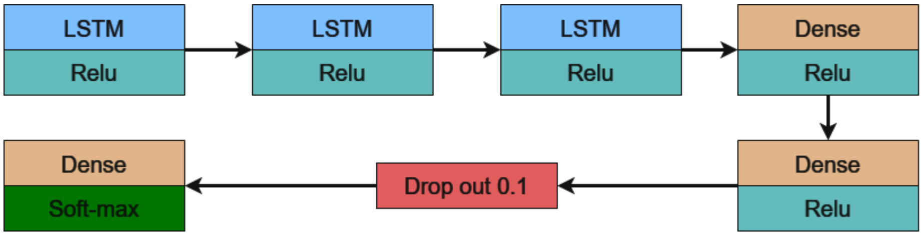

Fig. 4. Illustration of the model pipeline: LSTM Network

4 Experiment

We utilized the How2Sign dataset [5], a comprehensive collection of sign language videos, to train and test four independent models for sign language recognition: joint model, bone model, joint motion model, and bone motion model 3.3. We aimed to assess their performance and understand their strengths and weaknesses in recognizing sign language gestures. Each model underwent rigorous evaluation, and we compared their performance results to gain insights into their capabilities 4.3. Our research has implications for various domains, including assistive technologies, human-computer interaction, and education. Accurate sign language recognition systems can enhance accessibility for individuals with hearing impairments, allowing them to interact with technology and communicate more effectively. Furthermore, our findings contribute to the broader computer vision and pattern recognition field.

4.1 How2Sign Dataset

We utilized the How2Sign data introduced in the paper by Author et al. [5]. This data comprises Green Screen RGB clips (frontal view) and includes the following subsets:

Subsets	Words	Clips
Training	529,175	31,128
Validation	29,597	1,741
Testing	39,474	2,322
Total	598,246	35,191

Table 1. How2Sign Dataset Statistics

The How2Sign dataset [5] is valuable for sign language recognition research, providing a diverse collection of sign language gestures captured from multiple signers. The frontal view and green screen setup facilitate accurate pose estimation and reliable spatial information.

Researchers widely utilize the How2Sign dataset [5] to develop and evaluate sign language recognition models, benefiting applications in assistive technologies, human-computer interaction, and education. The dataset encourages advancements in deep learning, temporal modeling, and multi-modal fusion to enhance system accuracy and robustness.

In summary, the How2Sign dataset [5] plays a pivotal role in sign language recognition research, offering diverse gestures and supporting various applications. Its availability drives progress in the field, fostering inclusihttps://www.overleaf.com/project/64aff7f81bb8 communication and accessibility.

4.2 Training Method

We employed the MediaPipe [14] model to perform accurate pose estimation to enhance the training process. We utilized the obtained joint coordinates and converted them into numpy [22] files for efficient data reading and training. Each model was trained separately to accomplish different tasks, ensuring specialized training for specific recognition objectives.

For optimization, we adopted the Adam optimizer, a popular choice for training deep-learning models. Adam combines the advantages of adaptive learning rates and momentum, allowing for faster convergence and better generalization. This optimizer dynamically adjusts the learning rate for each parameter, leading to efficient model updates during training.

We employed the categorical cross-entropy loss function [23] to measure the performance and guide the training process. Categorical cross-entropy [23] is commonly used in multi-class classification tasks, such as sign language recognition, as it effectively captures the differences between predicted and actual class probabilities. Our models learn to make accurate predictions across various sign language gestures by minimizing this loss.

During training, we monitored the categorical accuracy [7], which measures the percentage of correctly classified samples. This metric serves as an evaluation criterion, reflecting the model's ability to recognize different sign language gestures correctly. By optimizing for both loss and accuracy, we aimed to develop models that minimize errors and achieve high precision in classification.

The architecture of our models, as depicted in the table below, includes LSTM [21] layers for sequence modeling, dense layers for feature extraction, and a dropout layer for regularization. The final dense layer outputs probabilities for each sign language class, enabling the models to make confident predictions. Through the combination of MediaPipe [14] pose estimation, appropriate optimization with Adam, and the utilization of categorical cross-entropy loss and categorical accuracy, our training method aimed to train models that effectively recognize sign language gestures with high precision and robustness.

Layer (type)	Output Shape	Param #
lstm_6 (LSTM)	(None, 30, 64)	442,112
lstm_7 (LSTM)	(None, 30, 128)	98,816
lstm_8 (LSTM)	(None, 64)	49,408
dense_6 (Dense)	(None, 64)	4,160
dense_7 (Dense)	(None, 32)	2,080
dropout_2 (Dropout)	(None, 32)	0
dense_8 (Dense)	(None, 3)	99

Table 2. Architecture of the Neural Network

4.3 Performance Results

When comparing the performance of the models on the validation set, we observed that the joint model achieved the highest performance among the four models. This finding supports the standard practice in sign language recognition of relying on joint coordinates for single models. Additionally, the ensemble model exhibited the highest overall performance. Our experiments determined the optimal alpha values to be 0.253, 0.252, 0.248, and 0.246 for the four models. The ensemble model outperformed the initial approach of using only the joint model, achieving an accuracy of 0.452.

Models	Accuracy	Precision	Recall	F1-score
Joint	67.04	68.12	67.04	65.32
Bone	66.81	67.89	66.81	65,1
Joint Motion	65.62	66.68	65.62	63.94
Bone Motion	65.25	66.30	65.25	63.58
Multi-Modal	67.34	68.43	67.34	65.62

Table 3. Performance of Different Streams

To assess the effectiveness of our proposed Multi-modal model, we conducted a comparative analysis with the Indian Sign Language Recognition (ISLR) model developed by Gupta et al. [8]. The ISLR model has garnered attention due to its notable performance in recent studies [8]. By comparing our Multi-modal model with this well-established approach, we aim to evaluate the advancements and uniqueness of our proposed solution in sign language recognition. This comparative analysis provides insights into our model's potential improvements and contributions.

Table 4.3 presents the performance comparison of the ISLR model and our Multi-modal model on Indian Sign Language recognition. We evaluate both models in terms of Accuracy, Precision, Recall, F1-score, and Average Execution Times. The results show that our Multi-modal model achieves an Accuracy of 0.858, slightly outperforming the ISLR model with an Accuracy of 0.855. Similarly, our model demonstrates competitive Precision, Recall, and F1-score values compared to the ISLR model. The Average Execution Times for both models are approximately 2 seconds.

Overall, the comparative analysis highlights the promising performance of our Multi-modal model in Indian Sign Language recognition, showcasing its effectiveness and potential advancements over the established ISLR model."

Models	Accuracy	Precision	Recall	F1-score	Average Execution Times
ISLR [8]	0.855	0.869	0.855	0.847	0.07s
Multi-modal 3.3	0.858	0.872	0.858	0.85	0.3s

Table 4. Performance of Different Models

5 Discussion

One of the essential metrics to evaluate the effectiveness of a sign language recognition model is accuracy. Accuracy is the ratio of the number of correct predictions and the total number of predictions. Accuracy is suitable for the problem of sign language recognition rather than using recall or precision because, in this problem, we care about both true negatives, which are the pauses between words or sentences. If we use recall or precision, we will ignore these pauses and only focus on true positives, which are the words or sentences that are correctly recognized. Thus, accuracy will tell us how accurate the model is in distinguishing between sign and non-sign language.

Categorical entropy loss is an ordinary loss function in multi-class classification problems, where the goal is to predict each input's label accurately. We need to classify videos or images into corresponding sign language symbols in sign language recognition. Categorical entropy loss can help us measure the difference between the model's probability distribution and the labels' actual distribution. By minimizing flat entropy loss, we can reduce the percentage of incorrect predictions of the model, thereby improving the effectiveness and accuracy of the problem. In addition, categorical entropy loss can help us detect and handle data issues, such as imbalance, noisy, or missing data, by adjusting weights or applying augmentation, imputation, or regularization techniques.

Our paper presents a method for sign language recognition using Long Short-Term Memory neural networks [21]. This method can recognize gestures and movements that form words, phrases, or sentences. This differs from previous approaches that only recognize single gestures, including letters and words. Our method has high accuracy with long gestures and movements but low accuracy with gestures that form letters and single words due to their similarity and other actions unrelated to sign language.

During our experimentation with sign language recognition using How2Sign [5] dataset, we observed that the Joint model consistently outperformed the Bone, Joint Motion, and Bone Motion models in terms of accuracy 4.3. This finding aligns with the common practice of incorporating joint coordinates or using them as the primary feature for sign language recognition [8]. The Joint model exhibited higher accuracy, highlighting the effectiveness of utilizing joint coordinates for processing sign language data. However, it is worth noting that the Multi-Modal model achieved the highest accuracy among all five models, suggesting the potential benefits of combining multiple modalities in sign language recognition.

6 Conclusion

6.1 Summary

This paper presented a sign language detection model using MediaPipe [14] with Long Short-Term Memory [21]. Our model can process video frames in real-time and recognize sign language gestures. The main advantage of our model is its speed, as it does not require complex preprocessing or feature extraction. The main disadvantage of our model is its accuracy, as it may fail to detect some subtle or complex gestures. We discussed possible ways to improve our model, such as adding more training data, fine-tuning the Long Short-Term Memory [21] parameters, and incorporating attention mechanisms. We hope that our model can contribute to developing sign language recognition systems and facilitate the communication of deaf and hard-of-hearing people.

6.2 Future works

In future work, we aim to develop a real-time translation system for deaf and hard-of-hearing individuals, enabling seamless communication with non-sign language users. This system would facilitate inclusive interactions and bridge the communication gap. Additionally, we propose enhancing the accessibility of online content by refining our sign language recognition method to automatically generate accurate subtitles or captions for videos, podcasts, lectures, and other multimedia formats. By providing adequate communication support, we can empower individuals with hearing impairments to engage with online resources and improve their overall accessibility fully.

Furthermore, we envision the creation of educational tools and interactive games that utilize sign language as a medium of instruction and interaction. These tools will promote sign language learning and cultural appreciation, catering to different proficiency levels and fostering broader adoption [6]. Future research should address challenges related to varying lighting conditions, hand orientations, and diverse signing styles to ensure practical implementation. Exploring multi-modal approaches 4.3 that incorporate facial expressions and emotions can enhance the accuracy and expressiveness of the SLR system. By pursuing these avenues, we aspire to develop innovative solutions that empower the deaf and hard-of-hearing community, improve accessibility, and promote the cultural significance of sign language.

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu) (2019)
2. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. NeurIPS (2022)
3. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.: Decoupling gcn with drop-graph module for skeleton-based action recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. pp. 536–553. Springer (2020)
4. Dardas, N.H., Georganas, N.D.: Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Transactions on Instrumentation and measurement **60**(11), 3592–3607 (2011)
5. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., i Nieto, X.G.: How2sign: A large-scale multimodal dataset for continuous american sign language (2021)

6. Emmorey, K.: Language, cognition, and the brain: Insights from sign language research. Psychology Press (2001)
7. Fan, X., Gaussier, É.: Supervised categorical metric learning with schatten p-norms. CoRR **abs/2002.11246** (2020), https://arxiv.org/abs/2002.11246
8. G, D.V., Goyal, K.: Indian sign language recognition using mediapipe holistic (2023)
9. Hu, L., Gao, L., Liu, Z., Feng, W.: Continuous sign language recognition with correlation network (2023)
10. Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y.: Skeleton aware multi-modal sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2021)
11. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 196–214. Springer (2020)
12. Köpüklü, O., Gunduz, A., Kose, N., Rigoll, G.: Real-time hand gesture detection and classification using convolutional neural networks. In: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019). pp. 1–8 (2019). https://doi.org/10.1109/FG.2019.8756576
13. Li, D., Opazo, C.R., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1448–1458 (2020). https://doi.org/10.1109/WACV45572.2020.9093512
14. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines (2019)
15. Pearce, T., Brintrup, A., Zhu, J.: Understanding softmax confidence and uncertainty. CoRR **abs/2106.04972** (2021), https://arxiv.org/abs/2106.04972
16. Schmidt, R.M.: Recurrent neural networks (rnns): A gentle introduction and overview. CoRR **abs/1912.05911** (2019), http://arxiv.org/abs/1912.05911
17. Shi, B., Rio, A.M.D., Keane, J., Michaux, J., Brentari, D., Shakhnarovich, G., Livescu, K.: American sign language fingerspelling recognition in the wild (2019)
18. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Transactions on Image Processing **29**, 9532–9545 (2020)
19. Sofianos, T., Sampieri, A., Franco, L., Galasso, F.: Space-time-separable graph convolutional network for pose forecasting. CoRR **abs/2110.04573** (2021), https://arxiv.org/abs/2110.04573
20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014), http://jmlr.org/papers/v15/srivastava14a.html
21. Staudemeyer, R.C., Morris, E.R.: Understanding lstm – a tutorial into long short-term memory recurrent neural networks (2019)
22. van der Walt, S., Colbert, S., Varoquaux, G.: The numpy array: A structure for efficient numerical computation. Computing in Science Engineering **13**, 22 – 30 (05 2011). https://doi.org/10.1109/MCSE.2011.37
23. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. CoRR **abs/1805.07836** (2018), http://arxiv.org/abs/1805.07836