Journal of Intelligent Systems and Control
https://www.acadlore.com/journals/JISC

Detection and Interpretation of Indian Sign Language Using LSTM Networks

Piyusha Vyavahare[1], Sanket Dhawale[1], Priyanka Takale[1], Vikrant Koli[1], Bhavana Kanawade[1*], Shraddha Khonde[2]

[1] Information Technology, International Institute of Information Technology, 411057 Pune, India
[2] Department of Computer Engineering, M. E. S. College of Engineering, S.P. Pune University, 411001 Pune, India

* Correspondence: Bhavana Kanawade (bhavanak@isquareit.edu.in)

Received: 06-10-2023 **Revised:** 07-05-2023 **Accepted:** 07-13-2023

Citation: P. Vyavahare, S. Dhawale, P. Takale, V. Koli, B. Kanawade, and S. Khonde, "Detection and interpretation of Indian Sign Language using LSTM networks," *J. Intell Syst. Control*, vol. 2, no. 3, pp. 132–142, 2023. https://doi.org/10.56578/jisc020302.

Abstract: Sign language plays a crucial role in communication for individuals with speech or hearing difficulties. However, the lack of a comprehensive Indian Sign Language (ISL) corpus impedes the development of text-to-ISL conversion systems. This study proposes a specific deep learning-based sign language detection system tailored specifically for Indian Sign Language (ISL). The proposed system utilizes Long Short-Term Memory (LSTM) networks to detect and recognize actions from dynamic ISL gestures captured in videos. Initially, the system employs computer vision algorithms to extract relevant features and representations from the input gestures. Subsequently, an LSTM-based deep learning architecture is employed to capture the temporal dependencies and patterns within the gestures. LSTM models excel in sequential data processing, making them well-suited for analyzing the dynamic nature of sign language gestures. To assess the effectiveness of the proposed system, extensive experimentation and evaluation were conducted. A customized dataset was curated, encompassing a diverse range of ISL sign language actions. This dataset was created by collecting video recordings of native ISL users performing various actions, ensuring comprehensive coverage of gestures and expressions. These videos were meticulously annotated and labelled with corresponding textual representations of the gestures. The dataset was then split into training and testing sets to train the LSTM-based model and evaluate its performance. The proposed system yielded promising results during the validation process, achieving a training accuracy of 96% and a test accuracy of 87% for ISL recognition. These results outperformed previous approaches in the field. The system's ability to effectively detect and recognize actions from dynamic ISL gestures, facilitated by the deep learning-based approach utilizing LSTM networks, demonstrates the potential for more accurate and robust sign language recognition systems. However, it is important to acknowledge the limitations of the system. Currently, the system's primary focus is on recognizing individual words rather than full sentences, indicating the need for further research to enhance sentence-level interpretations. Additionally, variations in lighting conditions, camera angles, and hand orientations can potentially impact the system's accuracy, particularly in the context of ISL.

Keywords: Indian Sign Language; Open CV; Deep learning models; LSTM

1 Introduction

Effective communication is vital for individuals to interact and thrive in society. However, individuals with limitations in speech or hearing face significant communication challenges. Sign language, as a visual language, plays a crucial role in enabling these individuals to express and understand communication through hand gestures, facial expressions, and body movements. Despite the importance of sign language, there remains a gap in effective communication between sign language users and those who do not understand sign language. This gap necessitates the development of sign language translation systems. This research paper aims to address the need for improved sign language translation by proposing a real-time system that leverages image processing and deep learning techniques. The primary objectives of this study are twofold: first, to develop a real-time sign language translation system, and second, to enhance the system's response time and accuracy in recognizing and translating sign language gestures. By

achieving these objectives, the research aims to contribute to the field of assistive technology, promoting inclusivity and improving communication accessibility for individuals with hearing impairments.

To establish the foundation for the proposed system, an extensive review of related work has been conducted. While previous research has explored sign language translation systems, limitations persist in terms of real-time translation, accurate gesture recognition, and accommodating the wide range of sign language variations and expressions. These research gaps and limitations motivate this study to explore novel approaches that can overcome these challenges and improve the effectiveness of sign language translation. The proposed methodology encompasses the utilization of image processing algorithms to analyze live sign motions and extract distinctive indicators, including hand gestures, facial expressions, and body movements. These indicators are then processed using classifiers to convert them into meaningful representations. An integral component of the proposed system is the incorporation of Long Short-Term Memory (LSTM) networks, which are renowned for their ability to capture long-term dependencies. By leveraging LSTM networks, the system aims to improve gesture recognition accuracy by effectively modeling the sequential nature of sign language expressions.

To train the system, a customized dataset focused primarily on Indian Sign Language (ISL) has been meticulously curated. This dataset encompasses a diverse range of sign language gestures, encompassing different hand shapes, orientations, movements, and positions. The inclusion of high-quality data in the dataset enables the system to learn and generalize from various sign language expressions effectively. The structure of the paper is as follows: Section 2 provides a comprehensive review of the related literature, highlighting the significance of gesture detection in sign language, discussing existing approaches, and identifying the research gaps and limitations in current methods. Section 3 presents the proposed methodology, outlining the image processing techniques, the integration of LSTM networks, and the details of the customized Indian Sign Language dataset. Section 4 describes the experimental setup and the evaluation of the system's performance. Finally, Section 5 concludes the paper, discussing the contributions of the research, potential future directions, and the impact on communication accessibility for individuals with hearing impairments.

2 Related Work

Each sign in a sign language is distinct due to changes in hand form, motion profile, and location of the hand, face, and other body parts, according to Rastogoo et al. [1]. It is challenging to identify visual sign languages using computer vision. The authors of the survey looked at the preceding five years' worth of deep learning- and vision-based models for sign language recognition. The proposed models demonstrate a notable increase in the taxonomy-based recognition accuracy of sign language.

A set of artificial intelligence technologies for sign language was proposed by Papastratis et al. [2]. The authors of the survey made an effort to provide a comprehensive overview of cutting-edge methods for sign language recording, recognition, translation, and representation while stressing both their advantages and disadvantages. The survey lists several uses while also outlining the major difficulties facing sign language technologies.

In the survey, deep learning models are used to identify and detect words in person's motion. Deep learning models are able to distinguish signals from single Indian Sign Language (ISL) video frames. They employ the feedback-based learning models LSTM and GRU. The four consecutive LSTM and GRU combinations (there are two layers of LSTM and two levels of GRU) were investigated by Kothadiya et al. [3] with their own dataset, IISL 2020. Over 11 different signals, the proposed algorithm, which consists of a single layer of LSTM followed by GRU, achieves approximately 97% accuracy.

Adaloglou et al. [4] reviews the most current developments in convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers for the recognition of sign language. The authors also identified the key challenges and open research questions in this field, such as the lack of standardized datasets and the need for real-time recognition in practical applications.

The authors proposed a two-stage recognition approach, which first detects the hand region in the input video frames and then recognizes the isolated hand sign from the detected hand region using a deep neural network. Rastgoo et al. [5] evaluated the proposed approach on two public datasets, namely, RWTH-BOSTON-50 and Polish Sign Language Alphabet (PSL). The paper concludes that the proposed deep cascaded model can be an effective solution for recognizing isolated hand signs in sign language and can be further extended to handle continuous sign language recognition.

Using large-scale video datasets like kinetics, Sarhan and Frintrop [6] optimize pre-trained action recognition models for sign language recognition tasks. The proposed method is evaluated on two publicly available sign language datasets, RWTH-PHOENIX-Weather and DGS-Corpus, and compared to state-of-the-art methods. The authors also analyzed the transferability of different pre-trained models and demonstrate that models trained on action recognition tasks that involve similar hand movements and poses as sign language can be effectively transferred for sign language recognition.

The system is designed to recognize vernacular sign languages, which are sign languages that are specific to a

particular region or community and may have limited resources for language recognition. Halder and Tayade [7] uses the Media pipe framework to extract hand and pose features from video input, and then apply machine learning algorithms, specifically Support Vector Machines (SVMs) and Random Forest (RF), for classification. The system is trained on a dataset of vernacular sign language gestures and evaluated on real-time video input. The results show that the proposed system achieves high accuracy in recognizing vernacular sign language gestures in real-time. The authors also compare the performance of SVM and RF classifiers and analyze the effect of different parameters on the system's performance.

Unique spatial and temporal information was extracted from each motion by Caliwag et al. [8]. The aspects that are gathered are then used to train a neural network to classify the gestures. Short, medium, and long motions were used in sign language videos as part of the movement-in-a-video recognition method. The author's suggested method had accuracy rates of 90.33% and 40%, compared to 69% and 43.7% for previous methods.

Continuous sign language recognition (SLR) tries to turn a series of signs into a complete thought. In the study, System presented the Structured Feature Network (SF-Net) to efficiently learn many levels of semantic information from the data and overcome these issues. After integrating the features, they obtained. Yang et al. [9] presented the SF-Net, which gradually encodes data at the frame, gloss, and sentence levels into the feature representation. The accuracy and adaptability of the SF-Net were evaluated using two sizable public SLR datasets collected from diverse continuous SLR scenarios. The results suggest that the SF-Net's accuracy and adaptability are superior to those of earlier sequence level supervision-based techniques.

Liao et al. [10] have introduced a new multimodal approach for dynamic sign language recognition, utilizing a system called BLSTM-3D residual network (B3D ResNet). This system combines bi-directional LSTM networks and a deep 3-dimensional residual ConvNet to automatically extract spatiotemporal features from video sequences and generate an intermediate score for each action in the video. The B3D ResNet system is composed of three main components: hand object localization, feature analysis, and video clip categorization. The researchers tested the system on two datasets, the DEVISIGN_D and SLR_Dataset, and obtained impressive recognition accuracy. Specifically, the B3D ResNet system achieved 86.9% accuracy on the SLR_Dataset and 89.8% accuracy. American Sign Language alphabet dataset was utilized by Thakur et al. [11] in their study. After being processed with Python libraries and tools such as OpenCV and skimage, the preprocessed gesture datasets are then trained using CNN VGG-16 mode. Since the technique explains the meaning of the hand signs used for conversing with someone who does not understand sign language, it allowed for one-way communication in the survey.

An innovative method was introduced by Starner et al. [12] with the title Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video System. Two real-time hidden Markov model-based systems were demonstrated, each of which tracks bare hands using a single-color camera in order to recognize continuous sentences of American Sign Language (ASL). The precision of the initial system, which utilizes a desk-mounted camera to monitor the user, is 92 percent, contrary to the level of detail of the other system, which makes use of a camera around the user's hat, which is 98%. For both experiments 40-word lexicon is used. It's worth noting that the development of real-time ASL recognition systems has great potential to improve communication. The use of wearable technology in this context is particularly promising, as it allows for more natural and unobtrusive interaction with technology.

The survey suggests a novel deep learning-based pipeline architecture that uses the Single Shot Detector (SSD), 2D Convolutional Neural Network (2DCNN), 3D Convolutional Neural Network (3DCNN), and Long Short-Term Memory (LSTM) to automatically recognise hand sign language. The hand skeleton is then constructed using the midpoint algorithm and the approximated key points. The scientists used 3DCNNs to extract discriminant local spatio-temporal characteristics from the stacked inputs, such as pixel-level, multi-view hand skeleton, and heatmap features. Authors discovered that 3DCNNs perform better than 2DCNNs at capturing the spatio-temporal dynamics of the hands after comparing their performances with various quantities of stacked inputs. Rastgoo et al. [13] employed a brand-new, sizable hand sign language dataset called RKS-PERSIANSIGN, which is made up of 10,000 RGB videos of 100 Persian sign words, to show their model. Another important contribution is the new RKS-PERSIANSIGN dataset, which offers a sizable real-world dataset for developing and testing hand sign language recognition algorithms.

To address the limitations of current hybrid HMM and CTC models for frame or word level alignment, Guo et al. [14] proposed a novel hierarchical-LSTM (HLSTM) encoder-decoder model with visual content and word embedding for Sign Language Translation (SLT). The HLSTM model is designed to handle different granularities, such as frames, clips, and viseme units, by capturing spatio-temporal transitions among them. After exploring the spatiotemporal cues of video clips using a 3D CNN, the authors packed the right visemes via online key clip mining. By combining the recurrent outputs of the top layer of the HLSTM, a temporal attention-aware weighting mechanism is used to balance the intrinsic link between the viseme source positions. Finally, viseme vector processing and semantic meaning translation are done separately using two LSTM layers. This new approach has the potential to improve the accuracy of Sign Language Translation by effectively capturing the spatio-temporal dynamics of

sign language. The HLSTM model with its visual content and word embedding enables the model to capture both the visual and semantic aspects of sign language, and the temporal attention-aware weighting mechanism further improves the model's ability to align and translate the spatio-temporal features.

Rastgoo et al. [15] proposed a novel model for recognizing isolated hand sign language using deep learning techniques from two input modalities, namely RGB and depth movies. The proposed model incorporates hand posture features and employs Scene Flow (SF) for depth video inputs and Optical Flow (OF) for flow information in RGB video inputs. The system employs a step-by-step investigation of various combinations of spatial information and several recurrent models, such as LSTM and GRU, to evaluate their impact on identification performance. The model achieved a competitive result on the isoGD dataset, which was only 0.22% lower than the most recent state-of-the-art model. This approach has the potential to enhance the accuracy of recognizing isolated hand sign language by effectively combining the strengths of different deep learning techniques and input modalities.

For the extraction of spatiotemporal features and sequence learning, Cui et al. [16] suggested recurrent convolutional neural network, to address the problem of glosses to video segments. The authors developed three stages of optimization using their proposed system architecture. The first stage involved building an end-to-end sequence learning framework with connectionist temporal classification (CTC) as the objective function, followed by suggesting an alignment. In the second stage, the alignment proposal was used as greater supervision to fine-tune the feature extractor. Finally, the authors presented the revised feature representations and optimized the sequence learning model. This approach has the potential to improve the accuracy of gloss-to-video segment alignment and enhance the performance of sign language recognition systems.

For weakly supervised continuous sign language recognition (SLR), Wei et al. [17] suggested a novel approach where only ordered bright tags are supplied for each sign video without temporal limits in frames. The suggested method uses a semantic boundary detection technique based on reinforcement learning to solve the problem of precisely aligning video frames with symbolic words in symbolic video. It formulates the problem of semantic boundary identification as a reinforcement learning problem and using a multilevel perceptual model to build discriminative representations for movies. The location of a semantic boundary is handled as an action, whereas the feature representation of a video segment functions as a state. Between the forecast statement and the ground truth statement, a quantitative performance measure is used to determine the reward. The policy gradient algorithm is used to train the policy network. On the CSL Split II and RWTH-PHOENIX-Weather 2014 datasets, the suggested technique has been thoroughly tested, and the results show its superiority and efficacy.

Word error rate (WER) is used as the primary assessment metric in a new continuous sign language recognition (SLR) architecture described by Pu et al. [18] that processes unaligned video-text pairs. Learning models are typically optimized by systematically removing connection time classification (CTC), as the WER is not separable. However, the WER metric may not always favor the phrase predicted with the highest decoding probability. The authors suggested a new architecture with increased diversity as a solution to this issue. They introduced multivariate data that replicated the WER calculation process, replacing, deleting, and adding the associated videos and text labels. They suggested multiple loss periods to reduce the gap between the video and the ground-truth label and to differentiate between real and apparent modes online using these actual and created pseudo-video-text pairings. Other CTC-based continuous SLR architectures that are already in use can easily be added to the framework. Extensive experimentation on two continuous SLR benchmarks, notably RWTH-PHOENIX-Weather and CSL, confirms the effectiveness of the presented technique.

A full convolutional network (FCN) for online sign language recognition (SLR) was created by Cheng et al. [19] and is capable of learning spatial and temporal information from poorly labelled video sequences with only sentence-level annotations. The authors added a GFE module to the suggested network in order to enhance learning about sequence alignment. Without any prior training, the network underwent end-to-end training. The results of the experiments demonstrated the effectiveness of the method and showed that it performed well in online recognition.

Sign language recognition has received a lot of attention in recent years because it can facilitate communication and increase access for the deaf. To address the challenges of low-resource sign language recognition, the Open Hands system, which integrates four crucial principles from the natural language processing (NLP) field to word-level sign language recognition, was introduced by Selvaraj et al. [20]. In order to provide implementation benchmarks and checkpoints, the system trained and published four pose-based isolated language recognition models in six languages, including American, Argentine, Chinese, Greek, Hindi, and Turkish. In addition, Open Hands published posture-based data and four different ISLR models in the same six sign languages. Evaluation of the open hands system showed that ST-GCN, a graph-based method, is accurate and effective in sign language recognition.

As per above discussion it is observed that there is a scope of research on database creation, preprocessing data, feature selection including the fact that the indicated existing technology could not manage the huge dataset with high accuracy. This issue is addressed in the proposed system using the LSTM neural network architecture. The strategy in the proposed system is based on one-way communication, and our dataset is built on dynamic ISL. The number of actions in our customized dataset is higher (40 actions) than in the current survey. Each action contains

30 videos and each video contain 30 frames. In this experimentation, Open CV Python library is used for gesture preprocessing. The system uses face, shoulder movements along with hand gestures.

3 Methodology

The study utilizes the Indian Sign Language dataset, which encompasses a diverse collection of 40 different actions performed using sign language. Each action is captured through 30 frames and 30 videos, ensuring a comprehensive representation. The dataset includes recordings from a varied group of subjects, ensuring representation across different genders and signing abilities. To capture precise hand movements and facial expressions, high-resolution cameras were used during the acquisition process. The subjects were instructed to perform the actions in controlled environments with adequate lighting and minimal distractions, and multiple takes were recorded to ensure data quality. However, it is important to acknowledge the limitations of the study. The dataset's size may restrict the representation of the entire spectrum of actions and variations observed in real-world sign language communication, potentially hindering the capture of diverse signing styles. Additionally, the subjects included may not fully represent the range of diversity found within the larger population of sign language users, potentially limiting the generalizability of the findings. The use of specialized equipment and controlled environments may fail to replicate the nuances of real-life sign language interactions, introducing biases related to factors such as lighting conditions, camera angles, and limited contextual information. Moreover, the subjective choices made during data collection, including action selection and framing, may further influence the dataset and impact the performance of the models. These limitations highlight the need for future research to address these issues by expanding the dataset, involving diverse subjects, incorporating more realistic environments, and adopting rigorous protocols to minimize bias and enhance the generalizability of the findings. Furthermore, the accuracy of the action recognition system relies heavily on the quality of the training data. Insufficient, biased, or incorrectly labeled training data can lead to poor performance and inaccurate predictions. Inaccurate detections or tracking, influenced by factors like lighting conditions, camera quality, occlusions, and variations in human poses, can also affect the accuracy of the Media pipe holistic model used for pose estimation and hand detection. Incorrect key point extraction resulting from inaccurate detections or tracking can consequently impact the action recognition results. Moreover, the performance and speed of the action recognition system are influenced by available hardware resources, including the CPU, GPU, and memory. Limited resources can result in increased processing time and impact the overall system performance.

3.1 System Architecture

Data processing: The first step as shown in the Figure 1 is to collect and process the data. This involves gathering video footage of people signing and labelling the footage with information about the gestures being made. The data is then pre-processed, cleaned, and transformed as necessary for use in the subsequent steps. Image Processing: The next step is to use computer vision algorithms to analyze the video footage and identify regions of interest, such as the hands and face of the signer. Image processing techniques such as image enhancement, segmentation, and feature extraction could then be used to extract relevant information about the hand gestures.

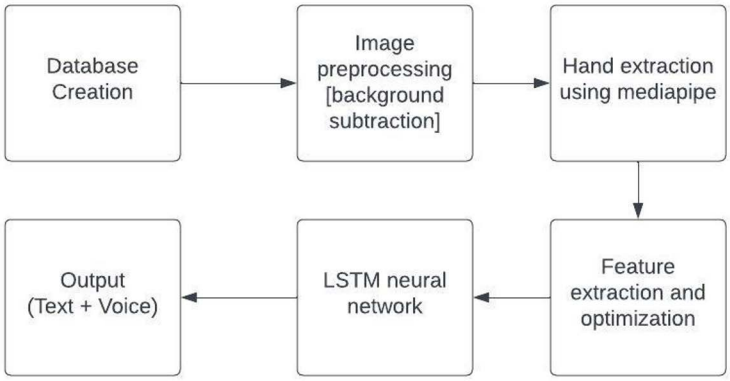

Figure 1. System architecture using LSTM

Body Gesture (Landmark Detection): Landmark detection techniques could be used to identify specific points on the hands, such as the fingertips or knuckles, to track the movement of the hands over time. This could involve using deep learning techniques such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to analyse the video frames and identify key landmarks.

Text + voice: Once the hand gestures have been identified, they can be translated into text or spoken language using natural language processing (NLP) techniques. To interpret the movements and provide the necessary output, one might need combining rule-based and machine learning-based approaches.

LSTM Neural Network: LSTM neural networks is used to learn patterns in the sequence of hand gestures over time. This involves training the network on a large dataset of sign language video footage and using backpropagation to update the weights of the network over multiple training epochs.

Feature Extraction: Finally, feature extraction techniques is used to extract relevant features from the input data and transform them into a format that can be used by the LSTM network. This might involve techniques such as principal component analysis (PCA) or wavelet analysis to make the incoming data lower dimensional and emphasise important features.

Overall, a sign language recognition system using action detection would involve a combination of data processing, image processing, body gesture analysis, text and voice recognition, LSTM neural networks, and feature extraction techniques to identify and interpret hand gestures in sign language.

3.2 System Design

The primary feature of this system is to translate the sign language into text. Initially, widely used gestures have been tracked to train the system. The system works at word-level translations. The captured images need to be pre-processed. The system modified the images captured and trained the LSTM model to classify the signals into labels.

3.2.1 User interfaces

The output screen will display the processed video stream, and the bottom of the video display window will show the words that are expected. Three words will be displayed, arranged from left to right in decreasing likelihood order from high to low. A vibrant outline will be used to highlight the term with the highest likelihood. Figures 2- 5 show what the actual interface will look like when Kinect completes the task. As shown in Figure 2, Kinect is carrying out a time-related action, and the findings are shown at the very top of the screen. Likewise, the results of the activities known as driving, saying sorry, and brushing are shown in Figure 3, Figure 4, and Figure 5.

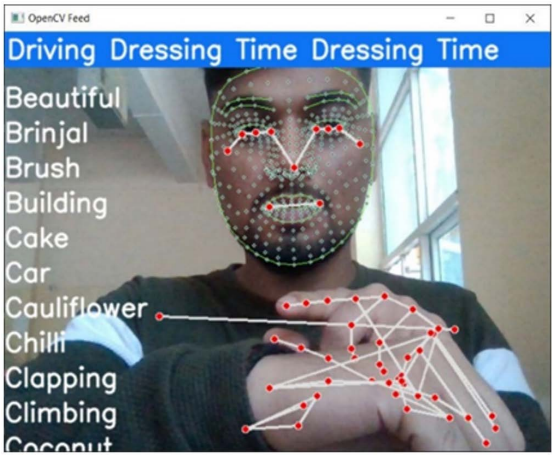

Figure 2. Action for time

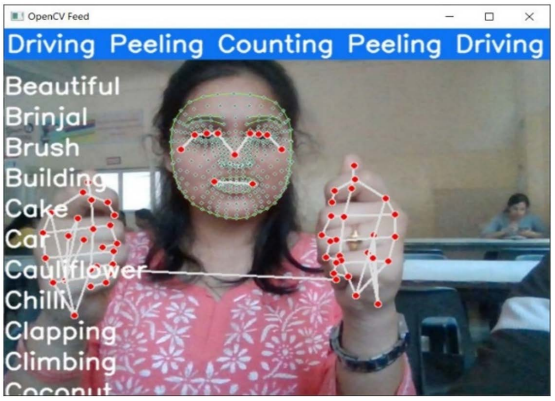

Figure 3. Action for driving

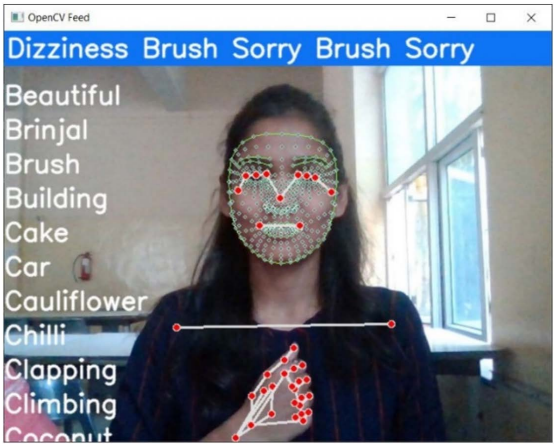

Figure 4. Action for sorry

Figure 5. Action for brush

3.2.2 Hardware interfaces

If the device lacks an integrated camera, an external camera sensor, as well as the driver required to enable the functionality on that specific operating system and hardware platform, will be required.

3.2.3 Software interfaces

OpenCV: It is used to capture the frame the input stream and then it is fed to the Media Pipe interface.

Media Pipe: It extracts the feature points tracked by OpenCV and the feeds it to the LSTM model for prediction.

TensorFlow is an open-source software library for high-performance numerical computing. Computing may be easily implemented across a variety of platforms (CPUs, GPUs, and TPUs), from desktop PCs to server clusters, due to its modular design.

3.3 Dataset

For testing and training, we produced the customized dataset. Both male and female participants made up the dataset, which was compiled with different subjects between the ages of 20 and 25. For specific dynamic-based isolated gestures, we were able to collect the appropriate data. The average frame rate (FPS) across all video samples is 40, and their average length is 4 seconds at a resolution of 1920 × 1080. 40 words made up the dataset. For each action 30 videos were taken, and each video contains 30 frames [21].

4 LSTM

Deep learning uses LSTMs, also known as long short-term memory networks. Numerous recurrent neural networks (RNNs) have the capacity to learn long-term dependencies, especially in tasks involving sequence prediction. In addition to processing single data points like visuals, LSTM features feedback links which allow it to process a whole data sequence. This is beneficial for speech recognition and machine translation, among other things. On an assortment of issues, the special RNN variant known as LSTM performs exceptionally well. The Long Short-Term

Memory (LSTM), an RNN structure that uses forgetting components, is usually suggested to address the gradient vanishing problem. One can select the optimal time delay and the appropriate period to forget certain information using this method. Given the properties of the known methods and the strategy suggested in this work, it is anticipated that a combination of these methods will have the ability to understand and distinguish sign language motions from a source of video and emit the corresponding English word. The LSTM model's structure is governed by a chain. However, the structure of a repeating module may change based on the application. The utilization of a single neural network was superseded by the creation an innovative system for communication featuring four interlinked levels. Figure 6 depicts the LSTM architecture. The input vector, the current memory block, the current output block, and the previous output block are all represented by the notation ht1, along with the memory block from the previous block.

Instead of having separate memory cells, GRU uses gating units to control the flow of information inside the unit. The mentioned method uses ISL gesture sequences in the video source that has been provided as an input. The major goal of this system is to recognize words from real-life signing movements. The first step is to divide the video file carrying the sequence of ISL motions for different words into individual sub-sections comprising different words. The proposed approach is a real-time system that uses image processing to process live sign gestures. The translated output would then display text and voice after classifiers were used to distinguish between different signs. To train on the data set, deep learning methods will be applied. With the use effective algorithms and high-quality data, its objective is to make the current system in this field more precise and quicker to respond. The potential of LSTM networks to learn long-term dependencies led to its study and implementation for the classification of gesture data. The developed model classifies the motions with high accuracy, demonstrating the viability of employing LSTM-based neural networks for sign language translation with the same parameters. After being received by the 1536-unit LSTM layer, data is delivered from the GRU layer using the same configurations, including 0.3 dropouts and l2 kernel regularized. The outcomes are sent to a dense, layer that is fully connected. With an effective value of 0.3, the result is then transferred to the dropout layer.

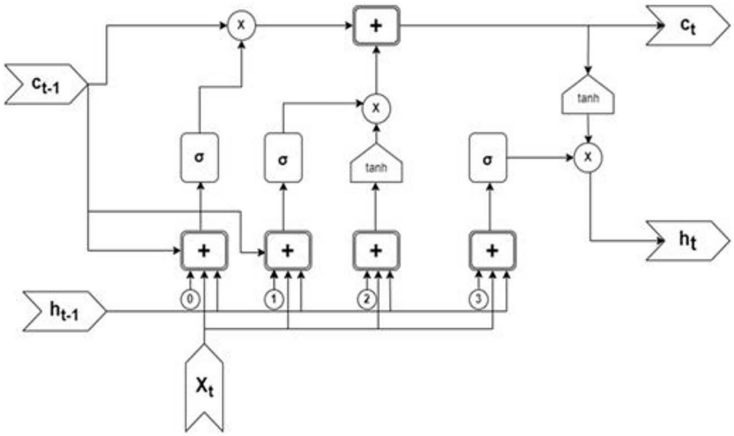

Figure 6. Architecture of LSTM

5 Results and Discussion

The analysis of the training loss and validation loss over the number of epochs provides valuable insights into the performance of the machine learning model. When both losses decrease together, it indicates that the model is effectively learning from the training data and generalizing well to new, unseen data. This is a positive sign as it suggests that the model is not overfitting the training data and has the potential to make accurate predictions on new data points. On the other hand, if the validation loss starts to rise while the training loss continues to decrease, it may indicate that the model is overfitting and is not able to generalize well. This information can guide us in making architectural considerations and necessary adjustments to improve the model's performance. Similarly, the relationship between training accuracy and validation accuracy over the epochs provides further insights into the model's capabilities. The fact that the training accuracy reached 87% indicates that the model is able to correctly classify a high proportion of the training data. However, it is important to note that the validation accuracy may differ from the training accuracy, as it measures the model's performance on unseen data. By examining the graph depicting the training accuracy and validation accuracy, we can observe whether the model is able to maintain a similar level of performance on new data or if there is a significant drop in accuracy. This understanding is crucial in determining the model's practical significance and its ability to make accurate predictions in real-world scenarios.

Overall, analyzing the training loss, validation loss, training accuracy, and validation accuracy provides us with valuable insights into the model's learning and generalization capabilities. These insights help us make informed decisions regarding the model's architecture, identify potential issues such as overfitting, and make necessary adjustments to improve its performance on unseen data. By continuously monitoring and evaluating these metrics, we can ensure that the model is effectively learning and producing reliable predictions, thus making it more practical and impactful in real-world applications.

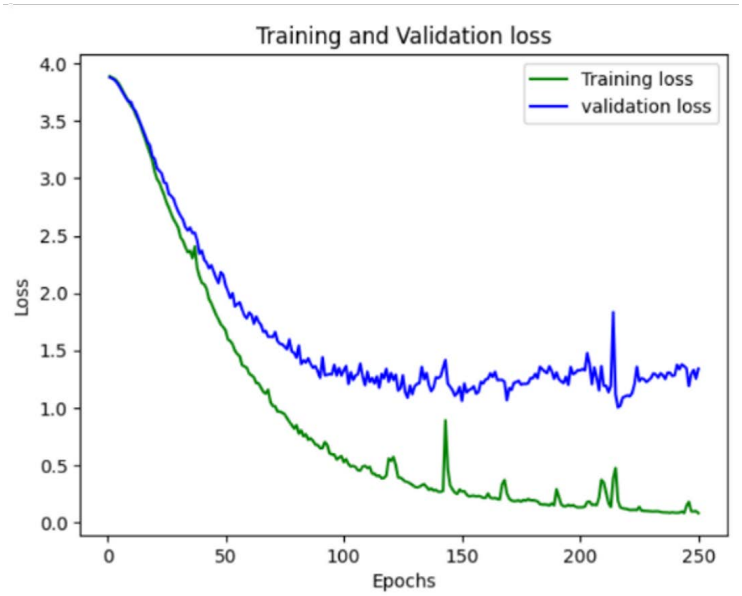

Figure 7. Training and validation loss

By comparing the training loss and validation loss over the number of epochs, we can get insights into the model's performance. If the training loss and validation loss decrease together, it's a good sign that the model is learning from the training data and generalizing well to new data The training loss vs. validation loss is displayed against the number of epochs in Figure 7 as a graph. This assisted us in making defensible judgements regarding the necessary architectural considerations. The graph in Figure 8 shows the relationship between training accuracy and validation accuracy over the number of epochs. The testing accuracy attained with references to this graph is 87%.

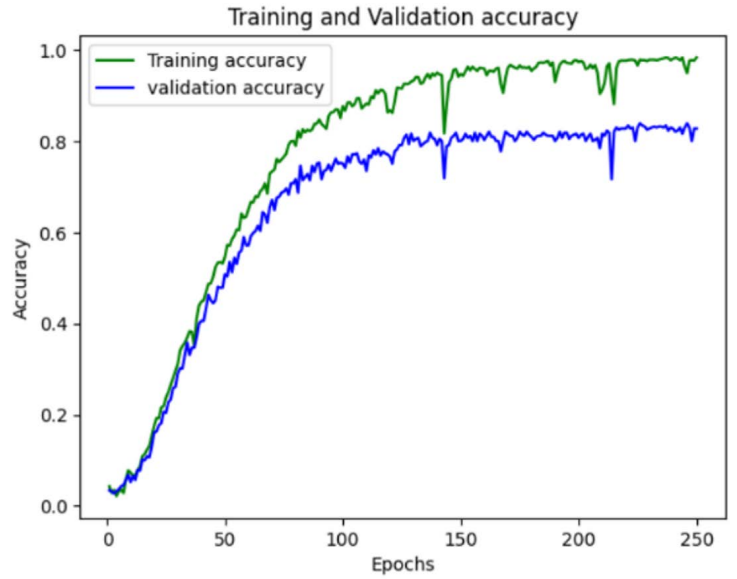

Figure 8. Training and validation accuracy

6 Conclusion and Future Scope

This study presents a technique for sign language action detection using LSTM on a proprietary dataset, with the objective of recognizing complete sign actions. The findings demonstrate that the proposed approach achieves a sign-action recognition accuracy up to 87%. However, it is important to acknowledge limitations stemming from model uncertainties, dataset biases, and experimental errors that may restrict generalization. To enhance the proposed approach, theoretical improvements can be made by exploring advanced neural network architectures and incorporating transfer learning. Practical recommendations include augmenting the dataset with more diverse examples, evaluating real-time scenarios, and engaging with sign language experts and communities to ensure accuracy and cultural sensitivity.

To build on the current study, future work can focus on several directions. Firstly, improving the image processing component can enhance the accuracy of hand and joint detection, enabling more precise sign language action recognition. Secondly, further advancements can be made in translating sequences of sign language movements into text and speech, refining the system's ability to facilitate bidirectional communication between sign language and spoken language users. Increasing the size and variability of the dataset would contribute to better generalization and robustness of the model. Additionally, validating the results with additional subjects and videos would provide a more comprehensive evaluation of the proposed technique. These future endeavors would strengthen the practical implications of the study by refining the accuracy, inclusivity, and accessibility of sign language recognition systems, ultimately enabling more effective and inclusive communication between sign language and spoken language users.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Exp. Syst. Appl.*, vol. 164, p. 113794, 2020. https://doi.org/10.1016/j.eswa.2020.113794

[2] I. Papastratis, C. Chatzikonstantinou, D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Artificial intelligence technologies for sign language," *Sensors*, vol. 21, no. 17, p. 5843, 2021. https://doi.org/10.3390/s21175843

[3] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A. Gil-González, and J. M. Corchado, "Deepsign: Sign language detection and recognition using deep learning," *Electronics*, vol. 11, no. 11, p. 1780, 2022. https://doi.org/10.3390/electronics11111780

[4] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 1750–1762, 2021. https://doi.org/10.48550/arXiv.2007.12530

[5] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimed. Tools Appl.*, vol. 79, pp. 22 965–22 987, 2020. https://doi.org/10.1007/s11042-020-09048-5

[6] N. Sarhan and S. Frintrop, "Transfer learning for videos: From action recognition to sign language recognition," in *2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates*, 2020, pp. 1811–1815. https://doi.org/10.1109/ICIP40778.2020.9191289

[7] A. Halder and A. Tayade, "Real-time vernacular sign language recognition using mediapipe and machine learning," *Int. J. Res. Publ. Rev.*, vol. 2, pp. 9–17, 2021.

[8] A. C. Caliwag, H. J. Hwang, S. H. Kim, and W. Lim, "Movement-in-a-video detection scheme for sign language gesture recognition using neural network," *Appl. Sci.*, vol. 12, no. 20, p. 10542, 2022.

[9] Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai, "Sf-net: Structured feature network for continuous sign language recognition," *arXiv preprint arXiv:1908.01341*, 2019. https://doi.org/10.48550/arXiv.1908.01341

[10] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with blstm-3d residual networks," *IEEE Access*, vol. 7, pp. 38 044–38 054, 2019. https://doi.org/10.1109/ACCESS.2019.2904749

[11] A. Thakur, P. Budhathoki, S. Upreti, S. Shrestha, and S. Shakya, "Real time sign language recognition and speech generation," *J. Innov. Image Process.*, vol. 2, no. 2, pp. 65–76, 2020. https://doi.org/10.36548/jiip.2020.2.001

[12] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, 1998. https://doi.org/10.1109/34.735811

[13] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Syst. with Appl.*, vol. 150, p. 113336, 2020. https://doi.org/10.1016/j.eswa.2020.113336

[14] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical lstm for sign language translation," in *Proceedings of the AAAI conference on artificial intelligence, New Orleans, Lousiana, USA*, vol. 32, no. 1, 2018, pp. 6845–6852.

[15] R. Rastgoo, K. Kiani, and S. Escalera, "Hand pose aware multimodal isolated sign language recognition," *Multimed. Tools Appl.*, vol. 80, pp. 127–163, 2021. https://doi.org/10.1007/s11042-020-09700-0

[16] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA*, 2017, pp. 7361–7369. https://doi.org/10.1109/CVPR.2017.175

[17] C. Wei, J. Zhao, W. Zhou, and H. Li, "Semantic boundary detection with reinforcement learning for continuous sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1138–1149, 2020. https://doi.org/10.1109/TCSVT.2020.2999384

[18] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1497–1505. https://doi.org/10.48550/arXiv.2010.05264

[19] K. L. Cheng, Z. Yang, Q. Chen, and Y. W. Tai, "Fully convolutional networks for continuous sign language recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK*, 2020, pp. 697–714. https://doi.org/10.48550/arXiv.2007.12402

[20] P. Selvaraj, G. Nc, P. Kumar, and M. Khapra, "Openhands: Making sign language recognition accessible with pose-based pretrained models across languages," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland*, 2022, pp. 2114–2133. http://dx.doi.org/10.18653/v1/2022.acl-long.150

[21] V. Koli, "Action recognition in sign language detection using lstm," 2023. https://github.com/Vikrantkoli5/ACTION-RECOGNITION-IN-SIGN-LANGUAGE-DETECTION-USING-LSTM