Mini Project Report

on

Sign Language Recognition

Submitted by

Balusu Bhanu Prakash : 20BDS012

Bommaragoni Karthik: 20BDS014

Kandikonda Pavan Sashank: 20BDS030

Peddisetty Venkata Sai Pranay: 20BDS038

Under the guidance of

Dr. Animesh Chaturvedi

Assistant Professor

Department of Data Science and Intelligent Systems

Dr. Vivekraj

Assistant Professor

Department of Computer Science

DEPARTMENT OF DATA SCIENCE AND INTELLIGENT SYSTEMS

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

19/05/2023

Contents

1

[1] contents are clickable

List of Figures

List of Tables

1 Abstract

Sign language recognition plays a crucial role in bridging communication barriers between the deaf or hard of hearing individuals and the hearing community. Here, we offer a comprehensive study on sign language recognition, focusing on the development and evaluation of effective algorithms and systems for automatic interpretation of sign language gestures. Two different datasets of sign language are analysed and utilised for training and evaluation. We've built each model for each dataset. We've performed classical machine learning techniques on one dataset whereas used CNN on the other one. There are some potential applications of sign language recognition in areas such as assistive technologies, education, and human-computer interaction.

2 Introduction

Sign language recognition serves as a vital medium of communication specifically designed for individuals who are deaf or hard of hearing. It plays a pivotal role in facilitating effective interaction within the deaf community, allowing individuals to express their thoughts, emotions, and ideas through visual gestures and movements. However, the lack of familiarity with sign language among hearing individuals often creates barriers in comprehending and interpreting these gestures, thus limiting effective communication between deaf and hearing individuals.

To bridge this communication gap and promote inclusivity, sign language recognition technology has emerged as a promising solution. The primary objective is to develop sophisticated systems capable of automatically interpreting and translating sign language gestures into written or spoken language, thereby enabling seamless communication between individuals with hearing impairments and those without. This technology aims to break down barriers and create a more inclusive society where effective communication knows no bounds.

The implementation of a sign language recognition system involves leveraging cutting-edge machine learning and computer vision techniques. With advancements in deep learning algorithms and image processing, models can be trained to accurately recognize and interpret a diverse range of sign language gestures. Machine learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are employed to learn and extract meaningful features from the visual input, enabling accurate recognition and interpretation of sign language gestures.

The impact of sign language recognition technology is profound and far-reaching. By equipping individuals who are deaf or hard of hearing with a reliable and effective means of communication, this innovative technology empowers them to participate more fully in various aspects of life. In the realm of education, sign language recognition facilitates active participation and interaction for students with hearing impairments, creating a more inclusive learning environment that fosters equal opportunities. Students can engage in discussions, ask questions, and express their ideas using sign language, with the system automatically converting their gestures into written or spoken language for the benefit of both deaf and hearing individuals.

Moreover, in professional settings, sign language recognition technology breaks down communication barriers, providing individuals with hearing impairments with increased employment opportunities. Effective communication with colleagues, clients, and customers becomes seamless, allowing deaf individuals to contribute their skills and talents to various industries. This technology not only enhances job prospects but also promotes workplace diversity and fosters a more inclusive work environment.

In addition to education and employment, sign language recognition technology fosters greater social interactions. By enabling fluent and accurate communication between deaf individuals and the wider community, it nurtures understanding, empathy, and inclusivity. Hearing individuals can better understand and appreciate the richness of sign language, facilitating meaningful connections and reducing the sense of isolation that individuals with hearing impairments may experience.

The successful implementation of sign language recognition technology relies on several key stages. The initial stage involves the collection of an extensive dataset of sign language gestures that encompasses a wide range of signs, variations, and contexts. This dataset serves as the foundation for training and evaluating the sign language recognition system, ensuring its robustness and generalization capabilities.

To improve the system's performance, the collected dataset undergoes preprocessing and augmentation techniques.. The augmented dataset provides the models with a more comprehensive representation of sign language gestures, enabling them to generalize better and perform accurately in various real-world scenarios.The subsequent stage focuses on the training of deep learning models such as Convolutional Neural Networks (CNNs) helps the models to be trained well on the preprocessed and augmented dataset, leveraging powerful optimization algorithms to learn the complex patterns and representations inherent in sign language gestures.and then evaluation of the system's performance using appropriate metrics.

3 Related Work

Several studies have been conducted on sign language recognition, aiming to develop accurate and robust systems for interpreting sign language gestures. One prominent approach in this field is the utilisation of computer vision techniques combined with machine learning algorithms.

3.1 Sign Language Recognition using Depth and Skeleton Data

Li et al. (2014) proposed a method for sign language recognition that utilised depth and skeleton data captured by a Kinect sensor. The study aimed to improve the accuracy of recognizing sign language gestures by leveraging additional depth and skeletal information. The methodology began with the acquisition of depth and skeleton data using a sensor, which provided precise 3D information about hand movements and positions. The depth data captured the spatial distribution of the signer's hands, while the skeleton data represented the joint angles and positions.

.

To extract meaningful features from the acquired data, the author employed the Histogram of Oriented Gradients (HOG) method. HOG calculated the local gradient orientations in different image regions, providing an effective representation of the sign language gestures. This feature extraction step helped to capture relevant patterns and characteristics of the hand movements. The next step involved classification using a Hidden Markov Model (HMM). HMMs are probabilistic models that can capture the temporal dynamics of sign language gestures. The extracted features from HOG were used as input to train the HMM, allowing it to learn the patterns and transitions between different signs.

The training phase consisted of building an HMM for each sign in the sign language vocabulary. The HMMs were trained using a dataset containing examples of different signs performed by multiple signers. This allowed the model to learn the variations and characteristics of each sign gesture. During the recognition phase, the input sign language gesture was captured using the sensor. The HOG features were extracted from the depth and skeleton data of the captured gesture. The trained HMMs were then utilised to recognize and classify the input gesture based

on the learned sign models.

Li et al. evaluated their approach on a sign language dataset and reported promising results. The incorporation of depth and skeleton data, along with the HOG feature extraction and HMM-based classification, contributed to improved accuracy in sign language recognition.

Overall, Li et al.'s study demonstrated the effectiveness of utilising depth and skeleton data for sign language recognition. The combination of HOG feature extraction and HMM-based classification proved to be a valuable approach in capturing the spatial and temporal characteristics of sign language gestures. This research contributed to the advancement of accurate and robust systems for interpreting sign language, fostering inclusivity and accessibility for the deaf and hard of hearing community.

3.2 Spatial Feature Extraction for Sign Language Recognition using CNNs

Cui et al. (2016) explored the use of Convolutional Neural Networks (CNNs) for sign language recognition, specifically focusing on extracting spatial features from sign language images. Their research aimed to enhance the accuracy and efficiency of sign language recognition systems by leveraging the power of deep learning algorithms. In their methodology, Cui et al. first collected a diverse dataset of sign language images, containing various gestures performed by multiple signers. Each image in the dataset represented a specific sign gesture, providing a rich and varied training set. The authors recognized the importance of a comprehensive and representative dataset for training robust recognition models.

To extract spatial features from the sign language images, Cui et al. employed CNNs. CNNs have shown great success in various computer vision tasks, and their ability to automatically learn hierarchical features made them well-suited for sign language recognition. By applying multiple convolutional and pooling layers, the network learned to extract discriminative spatial features

from the input images. Following the feature extraction stage, a Support Vector Machine (SVM) classifier was used to classify the sign language gestures. SVMs are powerful supervised learning algorithms that can effectively handle high-dimensional data and perform accurate classification. By training the SVM classifier on the extracted features, Cui et al. aimed to achieve reliable recognition performance.

Cui et al. evaluated their approach on different sign language datasets, comparing the results with other state-of-the-art methods. The experimental results demonstrated the effectiveness of their proposed methodology. The combination of CNNs for spatial feature extraction and SVMs for classification yielded promising results in sign language recognition tasks, achieving high accuracy rates and outperforming previous approaches.

Overall, Cui et al.'s study showcased the potential of CNNs in extracting spatial features from sign language images and the effectiveness of SVMs in classifying these features. By leveraging deep learning techniques, their approach contributed to the advancement of accurate and efficient sign language recognition systems. The research highlighted the importance of dataset quality, feature extraction, and classification algorithms in achieving robust performance in sign language recognition tasks.

3.3 Incorporating Depth Information with 3D Convolutional Neural Networks

Pfister et al. (2014) presented a research paper that focused on incorporating depth information with 3D Convolutional Neural Networks (CNNs) for sign language recognition. Their study aimed to capture both spatial and temporal features from depth sequences, enhancing the performance of sign language recognition systems.

The methodology employed by Pfister et al. began with the acquisition of depth sequences of sign language gestures. Depth sequences provide additional information about the three-dimensional

structure and motion of the signer's hands, enabling a more comprehensive representation of the sign gestures. The depth sequences were captured using depth sensors or similar devices. To extract spatial and temporal features from the depth sequences, Pfister et al. introduced a 3D Convolutional Neural Network architecture. Unlike traditional 2D CNNs that process individual frames independently, 3D CNNs take into account the temporal dimension by incorporating multiple consecutive frames. This allowed the model to capture both spatial and temporal dependencies within the depth sequences.

The trained 3D CNN model learned to extract discriminative features from the depth sequences, capturing both the static hand configurations and the dynamic temporal changes during sign language gestures. By incorporating depth information and considering the temporal aspects of sign language, Pfister et al. achieved state-of-the-art performance on a large-scale sign language dataset, demonstrating the effectiveness of their approach.

In conclusion, Pfister et al.'s research highlighted the importance of incorporating depth information and temporal features for accurate sign language recognition. By introducing a 3D CNN architecture, their approach effectively captured both the spatial and temporal characteristics of sign language gestures from depth sequences. This advancement contributed to the development of more robust and precise systems for interpreting sign language, offering potential benefits for the deaf and hard of hearing community.

3.4 Spatio-temporal Attention-based RNN Model for Sign Language Recognition

Li et al. (2018) proposed a spatio-temporal attention-based Recurrent Neural Network (RNN) model for sign language recognition. Their research focused on capturing the temporal dynamics and long-term dependencies among sequential frames of sign language gestures, enhancing the accuracy of recognition systems.

In their methodology, Li et al. utilised a dataset of sign language videos, where each video consisted of a sequence of frames depicting different sign gestures. The sequential nature of the data necessitated a model capable of capturing the temporal dynamics. RNNs are well-suited for such tasks, as they can model the dependencies among sequential inputs.

To address the challenge of effectively capturing long-term dependencies, Li et al. incorporated a spatio-temporal attention mechanism into their RNN model. This attention mechanism enabled the model to dynamically focus on informative frames while downplaying less relevant ones. By attending to important spatial and temporal cues, the model could effectively extract meaningful features from the sign language gestures. The attention-based RNN model was trained on the sign language video dataset, leveraging the temporal dependencies among the frames. The model learned to assign higher weights to frames containing critical information and lower weights to less informative frames, thereby improving the discriminative power of the recognition system.

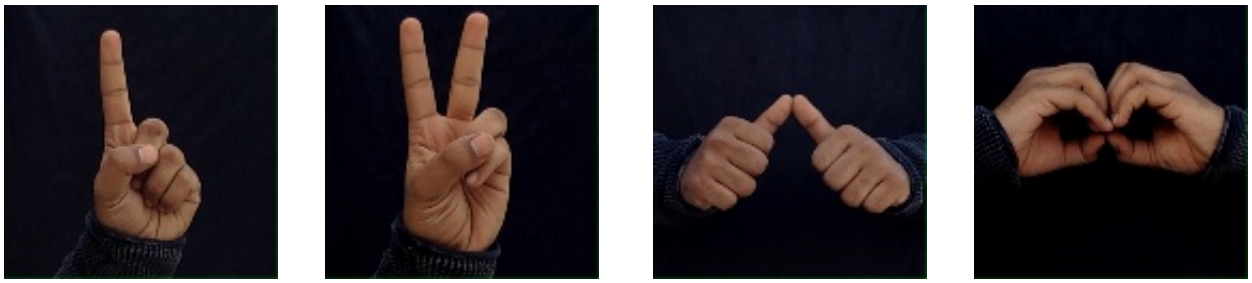
Li et al. evaluated their spatio-temporal attention-based RNN model on various sign language datasets and reported significant improvements in recognition accuracy compared to traditional RNN models. The attention mechanism allowed the model to effectively capture and utilise the temporal dependencies among sequential frames, leading to enhanced performance in sign language recognition tasks. In summary, Li et al.'s research demonstrated the effectiveness of incorporating spatio-temporal attention into RNN models for sign language recognition. By attending to relevant frames and capturing long-term dependencies, their model achieved improved accuracy in recognizing sign language gestures. This advancement contributes to the development of more sophisticated and precise systems for interpreting sign language, promoting inclusivity and accessibility for the deaf and hard of hearing community.

These previous works have demonstrated the potential of various techniques, such as HOG, CNNs, depth data, and RNNs, in the field of sign language recognition. By leveraging computer vision methods and machine learning algorithms, researchers have made significant progress in developing accurate and efficient systems for interpreting sign language gestures. These advancements helps them who rely on sign language as their primary mode of expression.

4 Datasets

4.1 ISL Dataset

The Indian Sign Language (ISL) dataset is a unique collection of data specifically created to facilitate Indian Sign Language recognition. It comprises a diverse range of images capturing various hand gestures and signs commonly used in ISL. These gestures encompass representations of alphabets, numbers, words, and phrases that are integral to Indian Sign Language communication. The dataset has been meticulously curated to serve as a valuable resource for the development and training of machine learning or computer vision models focused on accurately recognizing and comprehending these signs. Each sample within the dataset is thoughtfully labeled with the corresponding sign or gesture it represents, enabling algorithms to learn and effectively recognize the signs based on the provided training data. This comprehensive and well-organized dataset acts as a cornerstone for advancing sign language recognition technology in the context of Indian Sign Language.[1]



(a) 1 (b) 2 (c) A (d) B

Figure 1. ISL gestures

4.2 WLASL Dataset

The Word-Level American Sign Language (WLASL) dataset[2] consists of a large-scale collection of video clips, captured specifically to represent individual words in ASL. The dataset covers a broad vocabulary, incorporating words from various domains, including everyday objects, actions, emotions, and more.In total, we access 68,129 videos of 20,863 ASL glosses. Each video clip in the dataset is labeled with the corresponding word it represents, enabling supervised

learning and evaluation of word-level recognition algorithms.

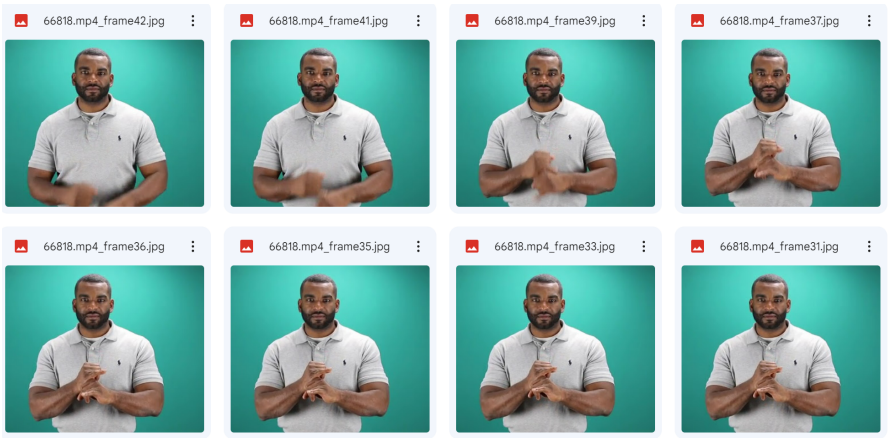


Figure 2. WLASL Dataset showing sign (year)

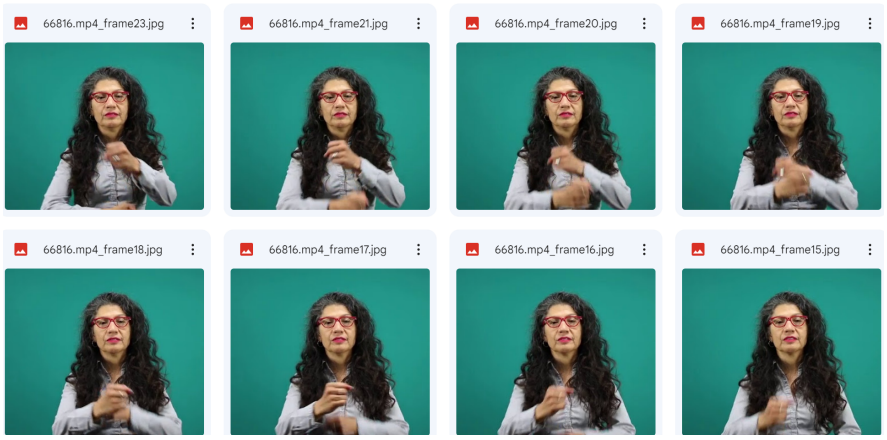


Figure 3. WLASL Dataset showing sign (year)

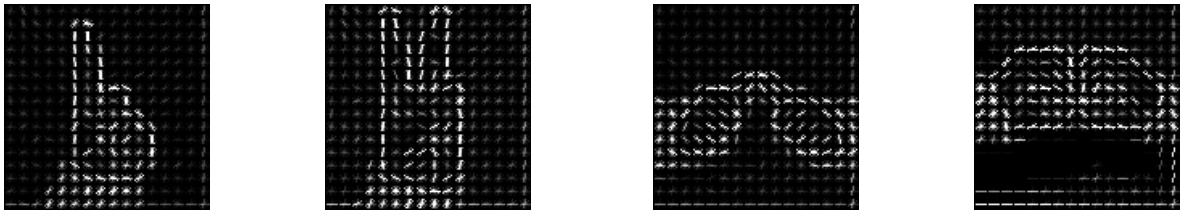
The dataset primarily comprises video clips in a standardized format, such as MP4, ensuring compatibility and ease of use. The videos are recorded in real-world scenarios, featuring different backgrounds, lighting conditions, and camera viewpoints. This variability reflects the diversity encountered during real sign language communication and enhances the robustness and generalization capabilities of trained models.

5 Techniques and Approaches

There are several methods to perform sign language recognition such as Deep learning techniques, such as recurrent neural networks (RNNs) or transformers, Sensor-based approaches that utilize wearable devices or motion sensors to capture and analyze hand movements. In glove-based systems, the user wears a specially designed glove embedded with sensors to track hand movements and gestures. Pose estimation techniques involve estimating the 3D pose or skeletal structure of the hand from 2D images or video sequences. Machine learning algorithms, such as convolutional neural networks (CNNs).In our application we have looked into Classical methods such as Support Vector Machine(SVM), Random forests, Decision tree, Naive Bayes, 2D convolution with Recurrnet Neural Networks(CNN) and Pose Based Temporal Graph Neural Networks(TGCN).

5.1 Classification Methods

In our application mod we try to investigate the classification of the Indian Sign Language (ISL) dataset using the Histogram of Oriented Gradients (HOG). The ISL dataset consists of video sequences capturing a variety of hand gestures utilized in ISL, with the objective of developing a classification model that is both accurate and efficient. By leveraging the HOG method, discriminative features are extracted from the hand gesture images, which effectively represent the local gradients of intensity. These extracted features are then given into a classifier, such as Support Vector Machines (SVM), Decision Trees, Random Forest, or Naive Bayes, to categorize the gestures into different classes. We try to assess and demonstrate the efficacy of the HOG-based approach for classifying ISL gestures, contributing to the advancement of ISL gesture classification techniques.[6]



(a) 1 (b) 2 (c) A (d) B

Figure 4. HOG images

5.2 2D convolution with Recurrnet Neural Networks

Convolutional Neural Networks (CNNs) have become widely adopted for extracting spatial features from input images, while Recurrent Neural Networks (RNNs) are utilized to capture long-term temporal dependencies in data sequences. To address this, our initial baseline model combines a CNN and an RNN to extract spatio-temporal features from video frames. Specifically, we employ the pretrained VGG16 model, which was trained on the ImageNet dataset, to extract spatial features. These features are then passed through a stacked GRU (Gated Recurrent Unit) network. This baseline model, referred to as 2D Conv RNN. To prevent overfitting on the training set, we set the hidden sizes of the GRU to 64, 96, 128, and 256 for the four subsets, respectively. Additionally, the GRU consists of 2 stacked recurrent layers. During training, we randomly select a maximum of 50 consecutive frames from each video and apply cross-entropy loss to both the output at each time step and the output feature obtained from average pooling of all the output features. During testing, we consider all frames in the video and make predictions based on the average pooling of the output features.[3][7]

5.3 Pose Based Temporal Graph Neural Networks

We propose a novel pose-based approach to Indian Sign Language Recognition (ISLR) using Temporal Graph Convolution Networks (TGCN). Our method focuses on the input pose sequence X1:N = [x1, x2, x3, ..., xN], representing concatenated 2D keypoint coordinates across N sequential frames. To capture spatial and temporal dependencies in the pose sequence, we introduce a graph network architecture. Unlike previous works that model motions using 2D joint angles, we encode temporal motion information by considering holistic trajectories of body keypoints.[5]

In our approach, we represent the human body as a fully-connected graph with K vertices, where the edges are represented by a weighted adjacency matrix $A \in \mathbb{R}^{K \times K}$. While a human body is only partially connected, constructing it as a fully-connected graph enables us to learn dependencies among joints through a graph network. In a deep graph convolutional network, each graph layer, denoted as G_n, takes a feature matrix $H_n \in \mathbb{R}^{K \times F}$ as input, where F represents the feature dimension from the previous layer. The initial layer takes the $K \times 2N$ matrix of body keypoint coordinates as input. By utilizing trainable weights $W_n \in \mathbb{R}^{F \times F_0}$ and a trainable

adjacency matrix A_n specific to the n-th layer, a graph convolutional layer is expressed as:

$$H_{n+1} = G_n(H_n) = \sigma(A_n H_n W_n) \quad (1)$$

where $\sigma(\cdot)$ denotes the activation function $\tanh(\cdot)$. To enhance information flow and facilitate learning, we utilize a residual graph convolutional block, which stacks two graph convolutional layers with a residual connection.

Our proposed approach leverages the Temporal Graph Convolution Networks (TGCN) to effectively model the spatial and temporal dependencies in the pose sequence for Indian Sign Language Recognition (ISLR). By representing the human body as a fully-connected graph and utilizing graph convolutional layers with trainable weights and adjacency matrices, we aim to capture the intricate relationships between body keypoints and encode meaningful motion information. The experimental results and evaluation of our approach demonstrate its potential in improving ISLR accuracy and performance.

5.3.1 Implementation details

The models, namely VGG-GRU, Pose-GRU, and Pose-TGCN, are implemented in the PyTorch framework. During training, we utilize the Adam optimizer. To ensure robust evaluation, we split the samples of each gloss into training, validation, and testing sets, maintaining a ratio of 4:1:1. We take care to include at least one sample per gloss in each split. The evaluation of the models is based on the mean scores of top-K classification accuracy, where K is set to 1, encompassing all sign instances. Figure 2 and Figure 3 illustrate that different meanings often exhibit similar sign gestures, which can lead to classification errors. However, these errors can be mitigated by considering contextual information. Therefore, it is more reasonable to employ top-K predicted labels for word-level sign language recognition tasks, improving the overall accuracy and performance of the system.[4]

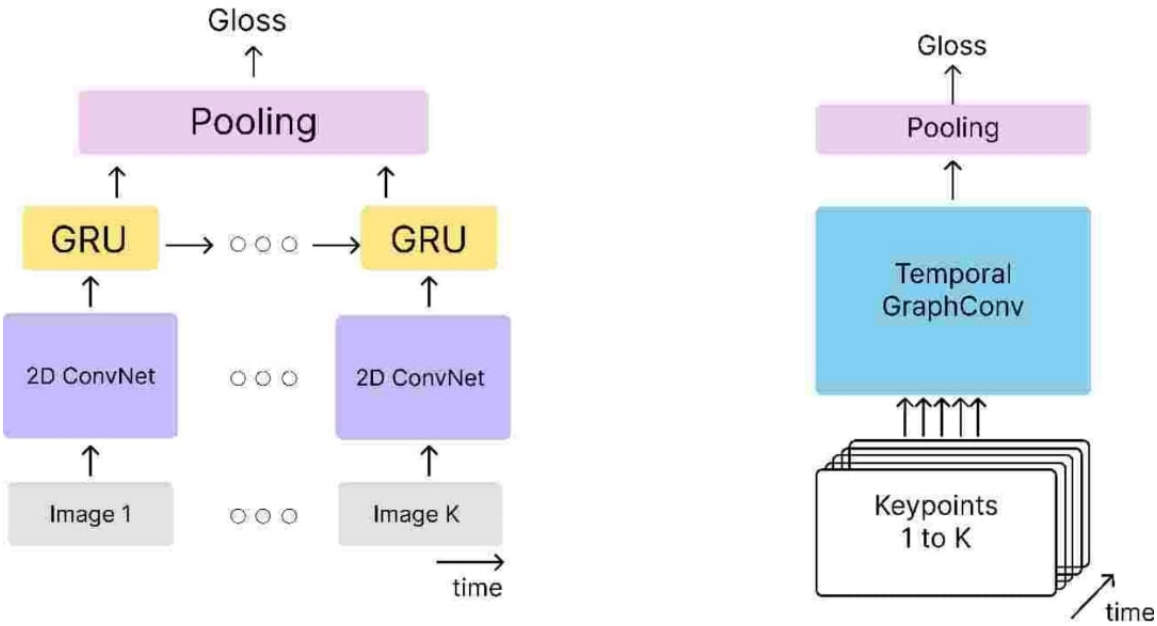


Figure 5. Our Baseline Architectures Fig a) 2D Conv. RNN
Fig b) Pose TGCN

6 Results and Discussions

To classify different inputs into the corresponding classes , we can use all of the standard classification techniques. The following are the results we've achieved from the methods we used.

Training Size	SVM	Decision Tree	Random Forest	Naive Bayes
50%	100%	98.67%	100%	99.6%
60%	100%	98.83%	100%	99.69%
70%	100%	99.15%	100%	99.74%
80%	100%	99.03%	100%	99.8%

Table 1
Accuracy table for classification models on ISL(Indian Sign Language) Dataset

It is quite evident that the classical methods are enough for the image level ISL dataset. Hence, Machine learning methods were not implemented on this dataset. Now looking into the WLASL dataset, we took the CNN approach with the below given parameters.

	CNN
Accuracy	96.80%
Validation Accuracy	98.04%
Batch Size	32
Optimizer	Adam
Epoch	7
Learning Rate	0.001

Table 2
Results for CNN on WLASL (Word-Level American Sign Language) Dataset

7 Conclusion

The primary objective of a sign language detection system is to provide an effective means of communication through hand gestures for both deaf and hearing individuals. The model demonstrates promising results, particularly in well-controlled lighting conditions and optimal intensity settings. To enhance the model's performance, additional gestures can be easily incorporated by capturing more images from different angles and frames. Scaling up the model to accommodate a larger dataset is also feasible, allowing for greater versatility. However, certain limitations exist, including reduced detection accuracy in challenging environmental conditions such as low light and uncontrolled backgrounds. Efforts will be made to address these issues and expand the dataset to achieve more precise and reliable results.

Why TGCN Model Failed?

- TGCN (Temporal Graph Convolutional Networks) is currently facing challenges due to runtime errors and high computational demands. The integration of heavy modules like cv2 (OpenCV) and others contributes to an increased overall weight, resulting in frequent runtime errors that hinder its smooth execution. Furthermore, TGCN's high computational requirements make it resource-intensive and lead to significant RAM usage.
- These modules are not specifically designed for the unique requirements of TGCN, making it prone to compatibility issues and inadequate memory allocation, resulting in runtime errors. The computational requirements of TGCN, combined with these heavy modules, lead to slow processing speeds and inefficient memory usage. This poses limitations on real-time applications and hampers scalability for large datasets or complex video sequences.
- To overcome these obstacles, it is crucial to develop lightweight and specialized modules optimized for TGCN's specific needs. By reducing reliance on heavy existing modules and adopting a streamlined approach, TGCN's potential can be unlocked, enhancing its performance and enabling broader applicability across different domains.

8 Future Work

The dataset can be easily expanded and customized to cater to the specific needs of users, representing a significant step in bridging the communication gap for individuals with speech and hearing impairments.

- Sign Language Research: Sign language recognition systems offer valuable tools for researchers investigating sign language linguistics, acquisition, and cultural aspects. These technologies facilitate large-scale data collection and analysis, fostering advancements in sign language research.

- The application of sign detection models in global meetings can greatly benefit disabled individuals by providing them with a means to understand and participate more effectively. The accessibility and user-friendliness of the model make it accessible to a wide range of users, including those with basic technological knowledge. Implementing such models at the elementary school level would enable young children to learn and appreciate sign language from an early age.

- Incorporating sign language recognition into assistive devices, such as smart gloves or wearable devices, empowers deaf individuals by enabling real-time translation of sign language into text or spoken language. This enhances their ability to navigate daily life, communicate effectively, and independently access various services.

- Communication aids, like sign language recognition systems, facilitate seamless translation between sign language and spoken language, fostering better understanding and interaction between individuals who are deaf or hard of hearing and the hearing community.

- Integrating sign language recognition into public spaces, including transportation systems, healthcare facilities, and government offices, reduces communication barriers. Deaf individuals can easily access information, receive instructions, and interact with service providers more efficiently, promoting inclusivity and equal access to services.

References

[1] Vaishnavi Asonawane. Isl dataset. https://www.kaggle.com/datasets/vaishnaviasonawane/indian-sign-language-dataset, 2020.

[2] Jessie Chatham Spencer Dongxu, Federico. Word-level american sign language (wlasl) dataset. GitHub repository, 2018. URL https://github.com/dxli94/WLASL.

[3] J. Huang, W. Zhou, and H. Li. Sign language recognition using 3d convolutional neural networks. 2021.

[4] Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. Task-oriented grasp affordance detection with convolutional neural networks. *arXiv preprint arXiv:1910.11006*, 2019.

[5] Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. Word-level deep sign language recognition. *arXiv preprint arXiv:2102.04575*, 2021. URL https://paperswithcode.com/paper/word-level-deep-sign-language-recognition.

[6] Navneet.Dalal, Bill.Triggs, 2005.

[7] Y. Ye, Y. Tian, M. Huenerfauth, and J. Liu. Recognizing american sign language gestures from within continuous videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.