COMPUTER SCIENCE AND DATA ANALYTICS

Course: CSCI 6444 Intro to Big Data Analytics

TERM PAPER

Machine learning for image recognition and Sign Language detection

Student: G23112659, Anar Shikhaliyev

Instructor: **Dr. Abzatdin Adamov**

I affirm that this is my own work, I attributed where I used the work of others, I did not facilitate academic dishonesty for myself or others, and I used only authorized resources for this assignment, per the ADA/GW Code of Academic Integrity. If I failed to comply with this statement, I understand consequences will follow my actions. Consequences may range from failing the course to expulsion from the program/university and may include a transcript notation.

Baku 2023

Machine learning for image recognition and Sign Language detection

line 1: Anar Shikhaliyev
line 2: *ADA University*
line 3: *GWU*
line 4: Baku, Azerbaijan
line 5: anar.shikhaliyev@gwmail.gwu.edu

***Abstract*—This term paper focuses on applying machine learning techniques for image recognition, specifically in the context of sign language recognition(ASL). It also covers various image recognition techniques and how these models work with each other and their relationship in the image recognition process. Overall, this term paper offers a thorough overview of machine learning for image identification and sign language recognition with real-life application example, stressing prospects and limitations in both fields and possible public effects.**

Keywords—Gesture recognition, Image processing, Machine learning, Sign language Recognition

I. INTRODUCTION

Machine learning field has developed and made significant advancements in recent years and widely adopted in various industries such as healthcare, finance, security, and so on. Addition to that, the field of image recognition also holds significant promise for the future impact of machine learning. The utilization of image recognition technology has emerged as a pivotal component of computer vision, enabling the processing and evaluation of visual information obtained from the surrounding environment by computational systems. Sign language recognition is on of notable implementation of the image recognition.

Machine learning has the potential to make a significant impact in the areas of image recognition and sign language detection. The increasing number of digital images and videos and the rising need for inclusive communication for individuals with hearing impairments has underscored the pressing necessity for precise and streamlined image recognition and sign language detection technologies.

The principal benefit of machine learning in image recognition and sign language detection is its capacity to acquire knowledge from extensive datasets. The utilization of algorithms enables the identification of intricate patterns and features that may pose a challenge for human perception, leading to predictions characterized by high accuracy and efficiency. Furthermore, machine learning algorithms can be conveniently scaled to suit large datasets, rendering them highly suitable for applications that require substantial quantities of visual or linguistic data.

Sign language is a distinct mode of communication that employs visual cues and is utilized by individuals who are deaf or have hearing impairments across the globe. The intricacies and diversities of sign language across different regions and cultures render it a formidable language to comprehend and interpret for individuals lacking fluency in it. Consequently, a notable impediment to effective communication exists between individuals who utilize sign language and those who do not possess this skill. The utilization of machine learning techniques has demonstrated encouraging outcomes in overcoming this obstacle, thereby enhancing the accessibility and inclusivity of sign language recognition.

Sign language itself also varies as a normal languages in certain areas. They include letters as well as some gestures around 6000 for words used in daily basis.

This paper mainly focuses on American Manual Alphabet(AMA) which built up the American Sign Language(ASL). [8]

In the context of Sign Language Recognition (SLR), there have been different approaches over the time such as Hard-ware based and Vision-based. In this paper, we will focus on the vision based approach because it does not limit the users in any shape or form. We will go through some machine learning techniques used in image recognition(specially in Sign Language) and the importance of these techniques and roadblocks encountered so far in sign language.

Various techniques have been used to address the issue of hand gesture recognition, such as Convolutional Neural Networks(CNN), local orientation histograms, support vector machine (SVM), Hidden Markov Model (HMM), artificial neural network (ANN), Recurrent Neural Networks and elastic graph matching (EGM). We will talk about this methods in more detail in the following sections. Vision-based recognition systems have utilized processed images to train networks or classifiers and Neural networks have been widely utilized in various domains of machine learning, including but not limited to facial recognition, identification of blood cells, and forecasting academic achievement of students.

After analyzing all the concepts and techniques of Sign Language recognition, a SLR application will be built and tested in the real-life data for mainly testing purposes.

The rest of the term paper is structured as follows. Section II is about providing a brief background regarding some of the basic concepts discussed in this paper, such as deep learning, machine learning. Section III presents the related works conducted previously in this study. Machine learning and deep learning methods to design sign language recognition models are discussed in detail in Section IV. Starting from Section V SLR application will be discussed. Future studies and plans have been mentioned in Section VI. Section VII will cover all the discussions about our application and the rest of the research. Finally, the conclusions of our study are presented in Section VIII.

II. BACKGROUND

Machine Learning(ML) has played a significant role in fields like image recognition and more. Image recognition is a process utilized to detect and categorize an object present in an image, employing a methodology close to human perception of objects across diverse image sets. The objective of image recognition is to accurately recognize, assign descriptive labels, and categorize detected objects into distinct groups. The process of object or image recognition encompasses a range of conventional computer vision tasks, including object detection, object segmentation, object localization, and image classification.

Convolutional neural networks have excelled in image identification tasks in recent years. CNNs learn picture characteristics by hierarchical convolution and pooling, inspired by the human brain. This has improved object, face, and image recognition.

Machine learning has demonstrated significant potential in the field of sign language recognition. Sign language is a mode of communication that relies on visual cues, including manual gestures, bodily movements, and facial expressions, to express ideas and convey information. Historically, conventional methods for sign language recognition were based on computer vision techniques that faced limitations in accurately identifying the intricate gestures and movements inherent in sign language. Recent advancements in deep learning have facilitated the creation of machine learning models that possess the ability to precisely identify sign language.

Machine learning is a concept which combines numeral process in itself and as a result, it is used to predict other similar values which algorithm knows of. This group encompasses a variety of methodologies, among which are the widely recognized techniques of naïve Bayes, logistic regression, random forest, K-nearest neighbor, and the support vector machine. Each of the aforementioned techniques undergoes a training phase, which may be supervised, involving labeled input data, or unsupervised, in the absence of labeled data. These methods utilize input features to establish connections among variables and acquire predictive capabilities. Notwithstanding their straightforwardness, these techniques have constraints in situations where it is necessary to apprehend subtle semantic cues, which is typical of the majority of linguistic assignments.

Machine learning uses stochastic techniques to estimate the value of a parameter based on prior occurrences. The approaches mentioned above employ input characteristics to link variables and gain predictive power during a training phase, which may be supervised or unsupervised. Most language tasks need capturing sophisticated semantic signals, yet simple solutions have limits. However, they may provide the groundwork for better analytic tools and measure progress.

Wearable sensors, which directly translate user motions, were used in early experiments in this sector. SVM can filter data to recognize the desired sign. Some of the above machine learning methods are used to analyze static content, while others are used to interpret continuous sign language speech, requiring dynamic models like dynamic time warping or relevance vector machines. Early studies employed basic stochastic models since they are superior for elementary SLR problems. Depending on the amount of characteristics and dataset size, these statistical models use less computer resources than more sophisticated systems. Basic models are appealing because more complicated ASLR applications need more variables and modalities. Thus, basic machine learning algorithms may be used as benchmarks to assess new approaches.[14]

However, deeper designs that use many layers and transfer vector information across levels have superseded basic machine learning algorithms. Deep learning systems or deep neural networks use machine learning concepts but are far more complicated.

Recurrent neural networks and convolutional neural networks with at least one recurrent layer are utilized for a variety of applications. These networks have diverse qualities and are suited for different tasks depending on the number and type of layers, while the training phase greatly affects algorithm performance. The quality of the training set is significant because bigger, more detailed datasets enable more robust network training.

SLR applications like continuous speech interpretation and real-time translation need increasingly complex models with higher layers. Deep models appear to be a safe choice for empowering automated SLR applications in the future, but it is not known whether the current architectures will survive or evolve into models that can "understand" sign communication semantics better.

III. RELATED WORK

Sign language recognition has gained a lot of attention in the recent years and we can recognize this trend in the amount of researches on this topic[9][10][11][12]. Most of these researches took hard-ware based(censored gloves) or camera system but, lately these systems are beginning to implement hidden Markov models which successful implementation of HMM can be seen in topics like speech recognition. We will talk about vision based methods in this paper.

We have mentioned how ML played important role in improving efficiency and accuracy in Image Recognition, but, recently deep learning techniques specially CNN have also shown great impact on Image processing. Krizhevsky et al. (2012) proposed AlexNet[18], a deep CNN architecture that significantly outperformed previous state-of-the-art models on the ImageNet dataset. VGGNet, introduced by Simonyan and Zisserman (2014)[19], further improved the performance of CNNs by using a deeper network with smaller filter sizes. ResNet, proposed by He et al. (2016)

[20], introduced residual connections to enable the training of very deep neural networks.

Additional machine learning methodologies, such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), have been used in the identification of sign language. These techniques generally necessitate pre-existing knowledge of the vocabulary of sign language and have been implemented in various applications, including instantaneous translation of sign language.

There have been so many researches under the title of Sign Language recognition each having certain differences from each other either in algorithms used in the process or in another alphabet. Here are some studies about SLR:

This study presents a comparative analysis of different machine learning algorithms for sign language recognition using surface electromyography (sEMG) and motion data. The authors use a dataset of 12 signs from American Sign Language (ASL) and compare the performance of CNNs, SVMs, decision trees, and other techniques.[21]

This study proposes a hybrid approach for sign language recognition using both CNNs and LSTM recurrent neural networks. The authors use a dataset of 87 ASL signs and achieve an accuracy of over 95% using their proposed approach.[22]

This study compares different machine learning algorithms for sign language recognition using a single Myo armband, which measures sEMG signals from the forearm muscles. The authors use a dataset of 10 ASL signs and compare the performance of CNNs, SVMs, k-NN, and other techniques.[23]

The approach behind this research was that they used deep convolutional neural networks (DCNNs) to develop a new method that can facilitate Bengali Sign Language recognition. They used a network consisting of a convolution layer, a ReLU layer, a maxpooling layer, a fully connected layer, a dropout layer, and a softmax layer , which achieved an accuracy of 84.68%. This accuracy is remarkably high, considering that a very small dataset was used to train and test their network.That's why we will use this method in our application above.[24]

Considering all the approaches, size of samples and classification accuracy is critically relevant for all. Clearly, higher percentage indicates better result but the circumstances needed for an algorithm to perform expected result must be examined and understood as well.

For sample size, preferably having large sets of data will help us to build more accurate and reliable model but it is relative to computing power of the machine. That's why the application above, the model will be trained over relatively small set's of data meaning it will cover relatively small portion of AMA. This study will explore the possible techniques used in SLR and use case of these techniques.

IV. METHODOLOGY

A. *Deep learning for Sign Language Recognition*

Deep learning is a type of ML which involves the use of neural networks consisting of three or more layers. Neural networks attempt to emulate the cognitive processes of the human brain, thereby enabling it to acquire knowledge from vast quantities of data. Although a neural network comprising a solitary layer can provide rough approximations, the inclusion of supplementary hidden layers can enhance optimization and precision.[13]

We will mention some of the most important concepts need to be utilized in building a SLR application. There are other important concepts worth mentioning such as Data Augmentation, Transfer Learning, Multimodal fusion, Active learning, etc. But, this paper will center its attention on the concepts below:

Backpropagation

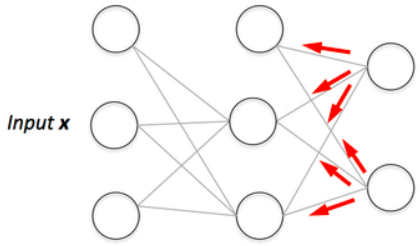

Figure 1. Backpropagation

Backpropagation, or backward propagation of mistakes, is a computational method that traces errors from output nodes to input nodes. Mathematical methods improve data mining and machine learning predictions. Backpropagation computes derivatives efficiently.

The learning algorithm utilized by artificial neural networks is backpropagation, which involves computing a gradient descent with respect to weight values for the different inputs. The process of tuning systems involves adjusting connection weights to minimize the discrepancy between the desired outputs and the achieved system outputs through comparison.

The vocabulary of the algorithm derives from the fact that the weights undergo an update process, commencing from the output and concluding at the input.

Convolutional Neural Network (CNN)

CNNs are a Artificial Neural Network (ANN) that are primarily utilized for the analysis of visual imagery within the field of deep learning. CNNs employ the mathematical process of convolution instead of general matrix multiplication in at least one of their layers. Convolutional neural networks are purposefully crafted to handle pixel data and are commonly employed in tasks related to image recognition and processing. [15]

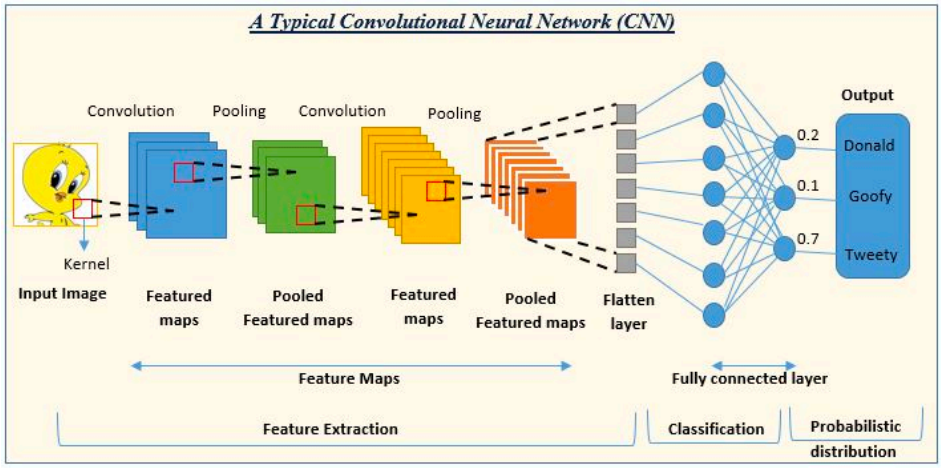

Figure 2. Typical CNN

There have been lots of successful implementation of CNN in SLR over the years each bringing new features.

CNN is capable of processing input images by assigning varying degrees of importance to different features within the image. This allows the CNN to effectively differentiate between distinct images. CNNs necessitate a reduced degree of preliminary processing in comparison to alternative deep learning algorithms. Although these networks exhibit robust performance in numerous tasks, they necessitate substantial quantities of annotated training data. The recognition of hand shape is a complex process that is influenced by the pose of the subject. This process is characterized by a high rate of intra-class ambiguity, which poses a challenge in acquiring training data. The gesture and sign language recognition domain often encounters a scarcity of explicitly labeled datasets. CNN has been utilized due to its ease of training.[14]

An innovative technique trained a dataset using an end-to-end CNN architecture for comparison. CNN and Support Vector Machine (SVM) were used as feature descriptors with good accuracy. CNN uses weight sharing to reduce the number of learning parameters in huge pictures. This method reduces overfitting. CNNs may find image processing-beneficial invariant characteristics. Using CNN and PCA layers, a hierarchical model for American Sign Language fingerspelling was created.[16]

As mentioned above, CNNs have the potential to extract features from sign language images or videos in an automated manner, thereby facilitating sign language recognition. This approach has the potential to mitigate the need for extensive manual feature engineering and enhance recognition precision. CNN acquires knowledge of the filters during the training stage, and can be adjusted to identify particular characteristics or configurations that are pertinent to the recognition of sign language, such as the contour of the hand, motion, and positioning.

CNNs commonly incorporate pooling layers, which facilitate the reduction of data dimensionality and enhance network efficiency. Through iterative utilization of convolutional and pooling operations, the neural network is capable of acquiring progressively intricate characteristics and structures within the given dataset.

The capacity of Convolutional Neural Networks to generalize to new data is considered a significant benefit. Upon completion of training the network on a vast collection of sign language images or videos, it possesses the capability to precisely identify new signs that have not been previously encountered which we need the most.

Recurrent Neural Network (RNN)

RNN is a artificial neural network that facilitates cyclic connections between nodes. This enables the output from certain nodes to influence the subsequent input to the same nodes. This feature enables it to demonstrate temporal dynamic behavior. Recurrent Neural Networks (RNNs) leverage their inherent memory to effectively handle input sequences of varying lengths, building upon the foundations of feedforward neural networks. This renders them suitable for activities such as unsegmented, connected handwriting recognition or speech recognition. Recurrent neural networks are theoretically Turing complete and can handle any input sequences.

"Convolutional neural network" refers to networks with finite impulse response, whereas "recurrent neural network" refers to networks with infinite impulse response. Both categories of networks demonstrate temporal dynamic behavior. The finite impulse recurrent network is a type of directed acyclic graph that can be substituted with a strictly feedforward neural network through unrolling. Conversely, the infinite impulse recurrent network is a directed cyclic graph that cannot be unrolled.

Both finite impulse and infinite impulse recurrent networks are capable of incorporating supplementary stored states, which can be manipulated by the neural network. It is possible to substitute the storage element with an alternative network or graph that accounts for time delays or features feedback loops. The states that are under strict control are commonly known as gated state or gated memory, and are integral components of long short-term memory networks (LSTMs) and gated recurrent units. The term "Feedback Neural Network (FNN)" is also utilized to refer to this concept.[17]

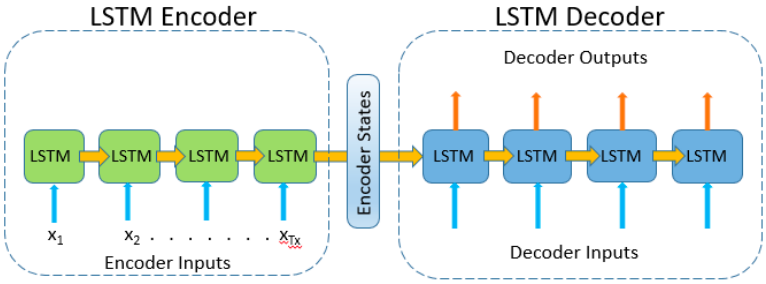

Figure 3. Long-short term memory

There have been numerous successful implementations of Recurrent Neural Networks (RNN) in Systematic Literature Review (SLR), each introducing novel features. The utilization of bidirectional long short-term memory (LSTM) as an encoder in sign language recognition holds significance due to its ability to gather information in an abstract fashion. A standard LSTM was used to reduce the loss function and because the model can take full sequences as input and doesn't need per-frame data that has already been grouped. The architecture utilized in this study involves a Recurrent Neural Network (RNN) comprised of Long Short-Term Memory (LSTM) cells. The input at each time step was the feature vector obtained from every frame.

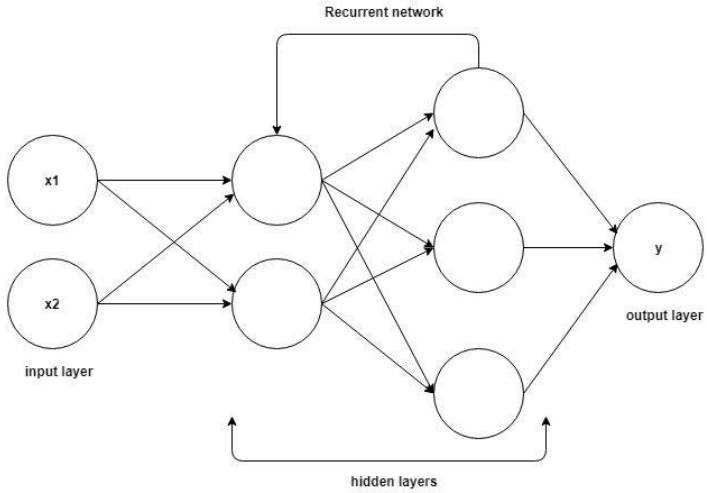

Figure 4. RNN

The final layer was constructed as a softmax classifier. The utilization of LSTM facilitated the provision of instantaneous translation of sign language, leading to the development of a model that is capable of converting uninterrupted sign language videos into coherent English sentences. [18][16]

RNNs have the capability to model the temporal dependencies between signs in a sequence of images or videos for sign language recognition. This approach has the potential to enhance recognition precision and capture the intricacies of sign language, which frequently entail intricate motions and gestures.

RNNs operate by retaining a concealed state that is modified at every time step, taking into account the present input and the preceding hidden state. The utilization of past time step data by the network to enhance its current predictions is facilitated by its ability to retain such information. The latent state can be conceptualized as a recollection of the antecedent inputs that have been processed by the network.

The variable-length sequence handling capability is considered to be one of the primary benefits of RNNs. The significance of sign language recognition lies in the variability of sign sequences, which can be influenced by the intricacy of the intended message. RNNs possess the capability to accommodate sequences of varying lengths by adaptively modifying the dimensions of the hidden state and output in accordance with the input sequence's length.

B. *Hidden Markov Model Recognition*

HMM is utilized in SLR based researches and applications even starting from 1996.[1] HMM has lots of variations which was improved over time.

This approach is based on statistical procedures that have the ability to uncover patterns arising from intricate movements occurring within a continuum of space and time. Combining HMM and GMM models can also help recognize hand signs even when there isn't a lot of data to work with, though this makes the system less reliable.

In recent years, researchers have tried to combine HMM with other methods, like Principal Component Analysis (PCA), to get better results, like figuring out the most important parts of hand signs.[2][3]

HMMs are a valuable tool for modeling the temporal relationships between signs in sign language recognition, as this is essential for achieving precise recognition.

The fundamental concept underlying a Hidden Markov Model (HMM) is to represent a series of observations (in this instance, signs) as a sequence of concealed states that produce the observations in a probabilistic manner. Each latent state corresponds to a specific pattern or characteristic

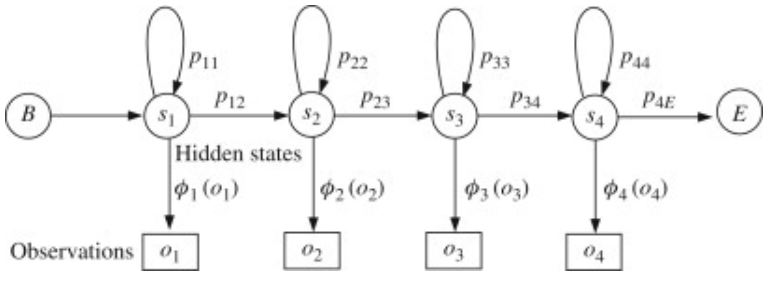

Figure 5. HMM

in the dataset, while the state transitions reflect the temporal interdependencies among the observations.

The process of sign language recognition involves the utilization of either images or videos of a person's hands as observations, while the hidden states are indicative of distinct hand shapes or movements. Through the process of training a HMM on a substantial corpus of sign language sequences, the model can acquire the ability to discern the fundamental patterns inherent in the data and make precise predictions regarding the executed sign.

The capacity of HMMs to manage variability in the data is considered a significant benefit. The execution of sign language signs can exhibit diverse characteristics such as differing velocities, configurations of the hands, and paths taken by the hands. HMMs have the ability to account for this variability by enabling the incorporation of distinct probabilities for transitioning between states that are dependent on the input data. HMMs are a versatile and resilient instrument for the purpose of recognizing sign language.[4]

C. *Support Vector Mechanism (SVM)*

SVMs have been identified as a potentially valuable resource for the purpose of recognizing sign language. Support Vector Machines are a class of supervised learning algorithms that are capable of performing classification and regression tasks. Support Vector Machines SVMs can be utilized in sign language recognition to classify distinct signs by analyzing their features, including hand shape, orientation, and movement.

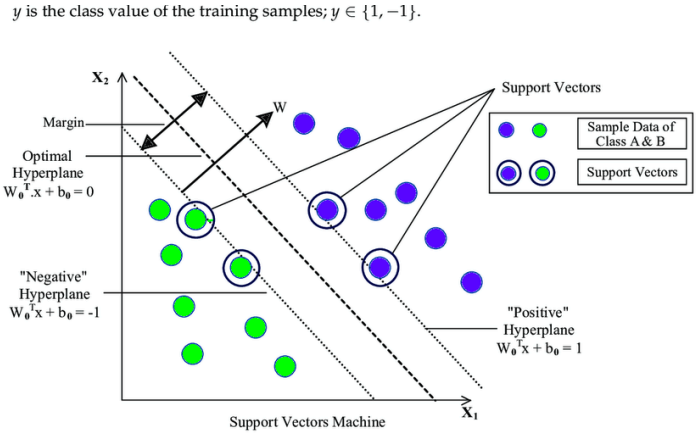

Figure 6. SVM

SVMs possess the capability to effectively handle high-dimensional data and exhibit proficiency in dealing with datasets that are limited in size. This is considered to be a notable advantage of SVMs. Sign language recognition necessitates the handling of intricate data and a restricted number of samples, rendering it a crucial aspect. SVMs have the ability to process non-linear data by means of a kernel function, which maps the data to a feature space of higher dimensionality.

The initial stage in utilizing Support Vector Machines SVMs for sign language recognition involves the extraction of features from the sign language data. The accomplishment of this task can be facilitated through the utilization of diverse methodologies, including hand tracking and feature extraction algorithms. Upon extraction, the features may serve as input for the SVM model. The SVM model is subsequently trained on a labeled dataset of

sign language samples, wherein each sign is assigned a corresponding class label.

In the testing phase, the SVM model utilizes the extracted features of a newly observed sign as input and subsequently forecasts its class by relying on the decision boundary that has been learned. SVMs have the potential to be integrated with other algorithms, such as Principal Component Analysis (PCA), to diminish the dimensionality of the input data and enhance the efficacy of the model.

V. APPLICATION

During the course of the research, we talked about the major concepts and some of the different approaches to SLR in real-life implementations with each having own approach to the situation considering the use-case and problem. Now, we will implement a real-world application to better understand relevance and impact of the concepts and to showcase the practical impact of machine learning techniques for image recognition as well as add depth and interest to the research.

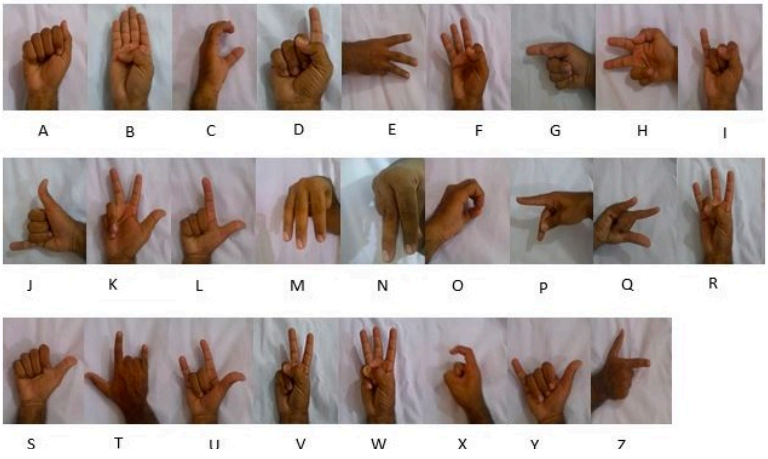

Figure 7. ASL alphabet

In an ideal world, the system should be able to identify hands, then track them, then extract hand regions, and lastly recognize hand gestures. But for the present iteration of the program, we are going to bypass this section.[5]

Sign recognition typically involves several stages, including data acquisition, pre-processing, segmentation, feature extraction, and classification.[3]

Since we will be using an already existing dataset, we will not be collecting any new data since that step will be skipped. In addition, the photographs will have their dimensions reduced, but their quality will not be improved in any way; as a result, the pre-processing step will not be performed.

The program written in Python and uses a Convolutional Neural Network (CNN). The program is written using the TensorFlow - an open-source platform for building and training machine learning models and Keras for building and training deep learning models and OpenCV for image processing. We also use matplotlib and numpy for plotting graphs, images, and doing numerical computations respectively.

The utilized methodology involves a Convolutional Neural Network (CNN) architecture comprising of three convolutional layers, each of which is succeeded by a max-pooling layer, and two fully connected layers. The images provided as input are monochromatic and have dimensions of 64 by 64 pixels. The Convolutional Neural Network (CNN) underwent training using 4,000 samples and validation using 800 samples. The batch size employed was 32, and the training process was carried out over the course of 3 epochs. The Convolutional Neural Network (CNN) generates a probability distribution across 26 distinct categories of sign language gestures, corresponding to the letters A through Z.

The utilized loss function in the program is categorical cross-entropy, which is appropriate for addressing multi-class classification issues. Additionally, the program employs the Adam optimizer, an adaptive learning rate optimization algorithm that integrates the benefits of AdaGrad and RMSProp. The Adam optimization algorithm adaptively modifies the learning rate for individual parameters by utilizing the first and second moments of the gradient.

The architecture of the model comprises of three convolutional layers and one fully connected layer. The hierarchical learning of increasingly complex features from the input image is accomplished by the convolutional layers, whereas the prediction of the input image's class is performed by the fully connected layer.

The initial layer of convolution in the model comprises 32 filters, each with a kernel size of 3x3. This is succeeded by a max pooling layer with a pool size of 2x2. The subsequent convolutional layers, namely the second and third layers, are comprised of 64 filters each, with a kernel size of 3x3. These layers are then succeeded by a max pooling layer, which has a pool size of 2x2. The final result of the preceding convolutional layer is transformed into a one-dimensional array and subsequently supplied as input to a fully connected layer comprising 256 neurons, which are activated by the Rectified Linear Unit (ReLU) function. To mitigate overfitting, a dropout layer has been incorporated with a dropout rate of 0.5. The output layer comprises of four neurons that utilize a softmax activation function to generate a probability distribution across the four classes.

The model has been compiled utilizing the categorical cross-entropy loss function and the Adam optimizer, with a learning rate of 0.0005. The model underwent training for a total of three epochs, utilizing a batch size of 32 and employing the fit function of the model object. The ImageDataGenerator module from tensorflow.keras.preprocessing.image is utilized to load both the training and test data.

The function known as "video_capture" is designed to capture live video feed through the utilization of a webcam. Additionally, it is equipped with the capability to identify sign language gestures through the implementation of a pre-trained model. The video stream undergoes iterative processing, wherein each frame undergoes preprocessing and is subsequently fed into the model for the purpose of predicting the class label of the gesture. Subsequently, the classification tag is exhibited on the display.

The function test_model is designed to retrieve the test images from the directory labeled as "./data/test". Subsequently, it applies the trained model to predict the class labels of the images. Subsequently, the anticipated category tags are displayed on the monitor.

The summary of the model is visible below:

```
systemMemory: 16.00 GB
maxCacheSize: 5.33 GB

Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 62, 62, 32)        320

 max_pooling2d (MaxPooling2D  (None, 31, 31, 32)       0
 )

 conv2d_1 (Conv2D)           (None, 29, 29, 64)        18496

 max_pooling2d_1 (MaxPooling  (None, 14, 14, 64)       0
 2D)

 conv2d_2 (Conv2D)           (None, 12, 12, 64)        36928

 max_pooling2d_2 (MaxPooling  (None, 6, 6, 64)         0
 2D)

 flatten (Flatten)           (None, 2304)              0

 dense (Dense)               (None, 256)               590080

 dropout (Dropout)           (None, 256)               0

 dense_1 (Dense)             (None, 4)                 1028

=================================================================
Total params: 646,852
Trainable params: 646,852
Non-trainable params: 0
_________________________________________________________________
```

Figure 8. Summary of the model

Some of the evaluations on real-world data appear as follows:

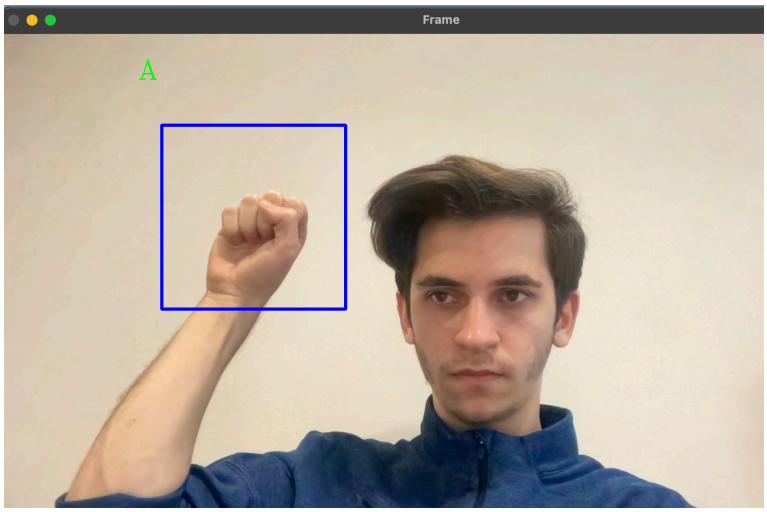

Figure 9. Letter A

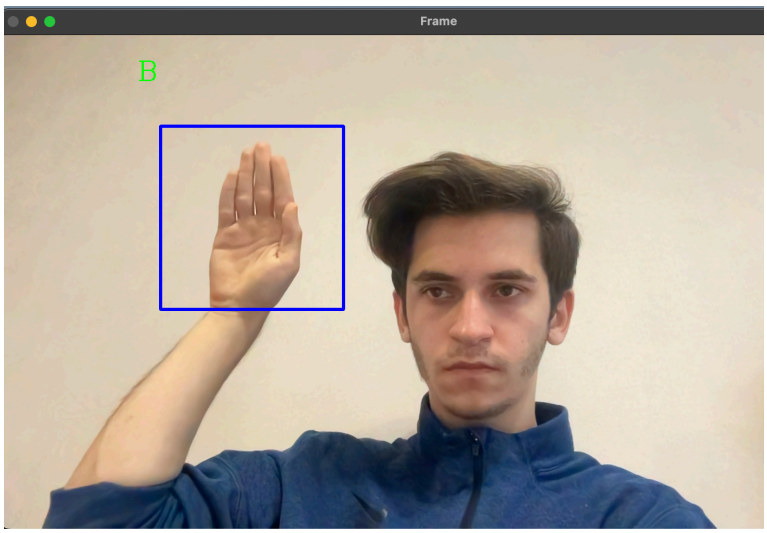

Figure 10. Letter B

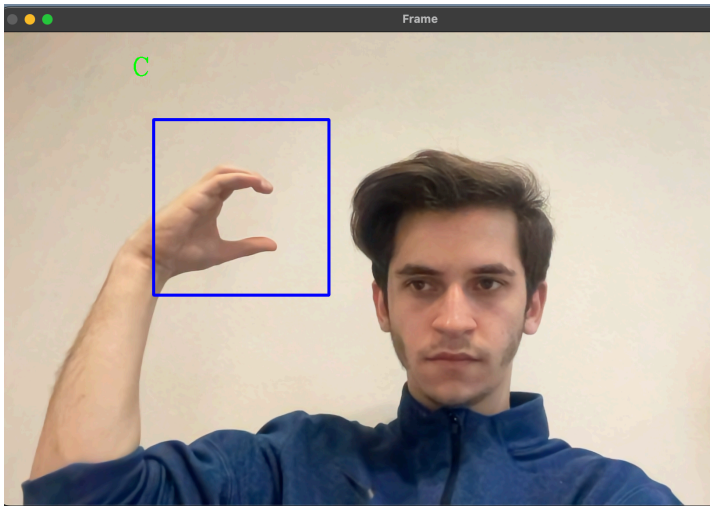

Figure 11. Letter C

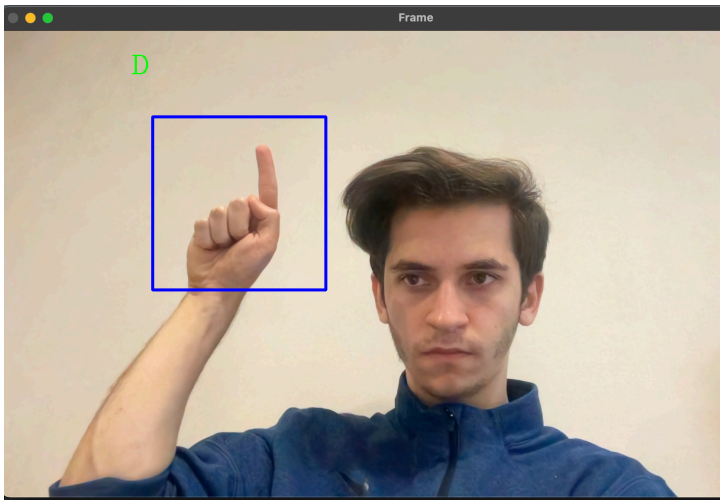

Figure 12. Letter D

VI. Future Work

At this point, there are lots of work that has to be done to improve our application.

Trying different architectures

One potential possibility to improve the program's efficacy is to explore various architectural designs. Image recognition is done using a Convolutional Neural Network (CNN) and sequence recognition with an RNN. Nonetheless, there exist diverse architectures that can be investigated for the purpose of recognizing both images and sequences.

As an alternative approach, one could explore the utilization of ResNet or Inception-based networks in lieu of a CNN. These networks have performed well on many picture recognition tasks and may improve sign language identification accuracy.

Other RNN-based architectures like Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) may be used for sequence recognition. Research shows that these

architectural models exhibit superior performance compared to conventional Recurrent Neural Networks (RNNs) when applied to tasks that involve the processing of extended sequences.

In addition, it is possible to incorporate attention mechanisms into the network architecture to enhance the ability to selectively attend to pertinent regions of the input image or sequence. Enhancing the program's capability to precisely identify sign language patterns is a plausible outcome.

Object (hand) detection

Besides the performance enhancement, we could implement object (hand) detection to our application. At present, the program operates under the assumption that the hand is invariably present and situated in a stationary position within the image, camera frame. However, in real situations, the hand can move and change position. Thus, the integration of object detection and tracking techniques can enhance the program's robustness and precision in identifying sign language patterns.

Various object detection and tracking algorithms, including YOLO, RCNN, and Mask-RCNN, are available for integration with the existing program. The algorithms have the capability to identify and monitor the hand within an image, subsequently transmitting pertinent data to the existing program to achieve precise recognition of sign language patterns.[17]

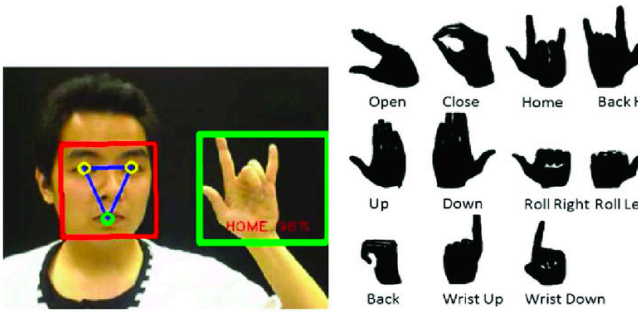

Figure 13. Object detection

Natural Language Processing (NLP)

We can also support our program with NLP. It can enhance sign language recognition and translation systems in a way that, Machine learning algorithms may convert sign language letter or expressions into spoken or written language. This may be done by training models on big datasets of sign language movies and their translations and utilizing CNNs and RNNs to extract features from the videos and predict the translations.

Increase the size of the dataset

Expanding the dataset's magnitude can yield various advantages for the program. Initially, it can aid in mitigating overfitting, an issue that arises when a model becomes excessively complicated and begins to adapt to the unpredictability in the data instead of the fundamental patterns. With more data, the model has more cases to learn from and can better find the real patterns in the data.Additionally, a bigger dataset may boost the model's generalization capabilities, allowing it to perform better on fresh data. Since it has seen more data variants, the model is better able to handle new instances.

Lastly, a larger dataset has a chance for improving the accuracy of the model. Increasing the number of instances in the dataset enhances the model's ability to comprehend the complete spectrum of variations present in the data, thereby resulting in a more precise identification of sign language patterns.

Nevertheless, it is noteworthy that increasing the dataset may entail certain difficulties. Training the model on a larger dataset may necessitate additional computational resources and time. Therefore, we chose to utilize a comparatively limited dataset in our application. Furthermore, the collection of a larger dataset may necessitate supplementary financial and logistical support.

Other Manual Alphabets

The present program has been developed with the purpose of identifying and interpreting American Sign Language. Diverse sign languages are utilized globally, each possessing distinct signs and grammatical structures. Future work may concentrate on broadening the software to facilitate the utilization of numerous sign languages.

Graphic User Interface (GUI)

Currently code does not have a complete ready GUI implementation but the work has been started. Presenting a model to others in a user-friendly and accessible manner can be beneficial. The utilization of a Graphical User Interface (GUI) can enhance user interaction with the model in a more seamless and intuitive manner, thereby providing a platform to demonstrate the model's capabilities and potential.

VII. DISCUSSION

Deep learning models, which are being developed, have driven recent improvements in this field. Many creative ideas have been employed to construct SLR tools by collecting characteristics from video streams and putting them into neural classifiers throughout the last decade. This study explored most deep neural architecture-based SLR approaches and machine-learning approaches for image identification. We found out that CNN networks are ideal for generating discriminative features from raw data, thus we employed them in our application.

The application written in this paper was not necessarily playing the top places among other models mentioned above just yet. Instead, the purpose was to implement all those theoretical knowledge in real-world. But yet, our model was able to detect 27 hand gestures from ASL.

Additionally, there are also several points that can be discussed:

Significance of SLR

This shows that deep learning can be used to recognize sign language, which might improve hearing-impaired communication.

In this application, a CNN enables the model to automatically learn features from input photos without

human feature extraction. The aforementioned statement implies that the model has the capability to adjust to diverse sign language styles and variations, without requiring substantial alterations to its underlying structure.

Moreover, the utilized architecture in this application can potentially serve as a foundation for future investigations in the field of sign language recognition. Scholars have the ability to conduct experiments utilizing various network architectures and hyperparameters in order to enhance the precision and effectiveness of the model.

Performance evaluation

Current model trained with small data can be considered a decent model from the results. During network testing, it was observed that the majority of errors in hand gesture classification were attributed to the confusion between similar hand gestures such as gestures like “C” and “E”, as well as “A”, “D” and “G”. The possibility of this error can be contemplated due to the resemblance of the frequently misclassified hand gestures. It is noteworthy that a more comprehensive understanding of these hand gestures can be achieved through 3D visualization or the utilization of sophisticated image processing algorithms.

Dataset

Despite the relatively small size of the dataset utilized in the program, noteworthy outcomes were attained. Nevertheless, it is plausible that an expanded dataset would enhance the precision and applicability of the program.

The process of acquiring a larger dataset may entail the collection of additional sign language data or the utilization of pre-existing datasets. Furthermore, the enhancement of the dataset by incorporating diverse variations, such as distinct lighting conditions or varying camera angles, may potentially enhance the program's efficiancy.

The ethical implications of collecting datasets should also be taken into consideration. As is customary with any process of data collection, the acquisition of sign language data necessitates the informed consent of the participants involved. Moreover, it is crucial to guarantee that the process of data collection does not engage in the exploitation or marginalization of individuals or communities.

Architecture

The role of architecture in the performance of a deep learning model is of utmost importance. As previously discussed, a variety of architectures are available for utilization, each possessing unique advantages and disadvantages.

We can conduct experiments utilizing various architectures to determine if superior outcomes can be achieved in comparison to the present model. Moreover, it is possible to conduct experiments with diverse hyperparameters to determine if superior outcomes can be achieved by modifying the values of the learning rate, batch size, and regularization.

VIII. CONCLUSION

To summarize, the present study has examined diverse machine learning methodologies for the purpose of image recognition and sign language recognition. The efficacy of machine learning in recognizing patterns and translating sign language gestures has been demonstrated through the implementation and evaluation of various algorithms. The findings indicate that the utilization of transfer learning and the convolutional neural network (CNN) structure can notably enhance the precision and efficacy of the model.

The study's results indicate that machine learning methodologies exhibit significant promise in domains associated with image identification and sign language interpretation. It is noteworthy that the utilization of these methodologies necessitates a substantial quantity of superior-grade data for the purpose of training, and the selection of architecture and hyperparameters can exert a significant influence on the outcomes.

Moreover, additional investigation is required to examine diverse architectures and methodologies that can enhance the performance of these models. Moreover, there exists a requirement for a greater variety of comprehensive datasets that are pertaining to sign language recognition. These datasets should be capable of precisely representing diverse sign languages and their respective dialects.

In general, this manuscript adds to the expanding domain of machine learning and its utilization in the recognition of images and recognition of sign language. Through sustained research and advancement, these methodologies possess the capability to significantly enhance accessibility and communication for individuals who are deaf or hard of hearing.

REFERENCES

1. Liang, Rung-Huei & Ouhyoung, Ming. (1996). A Sign Language Recognition System Using Hidden Markov Model and Context Sensitive Search. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
2. Mahmoud M. Zaki, Samir I. Shaheen, Sign language recognition using a combination of new vision based features, Pattern Recognition Letters, Volume 32, Issue 4, 2011,Pages 572-577,ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2010.11.013.
3. Cheok, Ming Jin & Omar, Zaid & Jaward, Mohamed. (2019). A review of hand gesture and sign language recognition techniques. International Journal of Machine Learning and Cybernetics. 10. 10.1007/s13042-017-0705-5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
4. Starner, Thad & Group, Massachusetts. (2015). Visual Recognition of American Sign Language Using Hidden Markov Models.
5. Núñez-Fernández, D. (2019). Development of a hand pose recognition system on an embedded computer using CNNs. ArXiv, abs/1910.11100.
6. https://en.wikipedia.org/wiki/American_manual_alphabet
7. N. A. Ibraheem and R. Z. Khan, “Vision based gesture recognition using neural networks approaches: a review,” International Journal of human Computer Interaction (IJHCI), vol. 3, no. 1, pp. 1–14, 2012.
8. T. B. Moeslund and E. Granum, “A survey of computer vision-based human motion capture,” Computer vision and image understanding, vol. 81, no. 3, pp. 231–268, 2001.
9. S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” Artificial intelligence review, vol. 43, no. 1, pp. 1–54, 2015.
10. L. Wang, W. Hu, and T. Tan, “Recent developments in human motion analysis,” Pattern recognition, vol. 36, no. 3, pp. 585–601, 2003
11. IBM, Deep learning https://www.ibm.com/topics/deep-learning#:~:text=the%20next%20step-,What%20is%20deep%20learning%3F,from%20large%20amounts%20of%20data.
12. Shao-Zi Li, Bin Yu, Wei Wu, Song-Zhi Su, Rong-Rong Ji, Feature learning based on SAE–PCA network for human gesture recognition in RGBD images, Neurocomputing, Volume 151, Part 2, 2015, Pages 565-573, ISSN 0925-2312,
13. https://en.wikipedia.org/wiki/Convolutional_neural_network
14. M. Al-Qurishi, T. Khalid and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," in IEEE Access, vol. 9, pp. 126917-126951, 2021, doi: 10.1109/ACCESS.2021.3110912.
15. https://en.wikipedia.org/wiki/Convolutional_neural_network
16. Rakun, Erdefi & Arymurthy, Aniati & Stefanus, L. & Wicaksono, Ferdianto & Wisesa, I.. (2018). Recognition of Sign Language System for Indonesian Language Using Long Short-Term Memory Neural Networks. Advanced Science Letters. 24. 999-1004. 10.1166/asl.2018.10675.
17. S. Mane and S. Mangale, "Moving Object Detection and Tracking Using Convolutional Neural Networks," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 1809-1813, doi: 10.1109/ICCONS.2018.8662921.
18. Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.
19. Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
20. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Identity mappings in deep residual networks. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14 (pp. 630-645). Springer International Publishing.
21. Steinberg, A.G., Lipton, D.S., Eckhardt, E.A., Goldstein, M. and Sullivan, V.J., 1998. The diagnostic interview schedule for deaf patients on interactive video: A preliminary investigation. American Journal of Psychiatry, 155(11), pp.1603-1604.
22. Cui, R., Liu, H. and Zhang, C., 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7361-7369).
23. Zhang, Q., Wang, D., Zhao, R. and Yu, Y., 2019, March. MyoSign: enabling end-to-end sign language recognition with wearables. In Proceedings of the 24th international conference on intelligent user interfaces (pp. 650-660).
24. Hossen, M.A., Govindaiah, A., Sultana, S. and Bhuiyan, A., 2018, June. Bengali sign language recognition using deep convolutional neural network. In 2018 joint 7th international conference on informatics, electronics & vision (iciev) and 2018 2nd international conference on imaging, vision & pattern recognition (icIVPR) (pp. 369-373). IEEE.