Sign Language Recognition

1st Harshita Khubchandani
Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
hk1319@srmist.edu.in

2nd Karthick T
Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
karthict@srmist.edu.in

2023 9th International Conference on Information Technology Trends (ITT) | 979-8-3503-2750-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITT59889.2023.10184252

***Abstract*— Deaf and hard-of-hearing persons communicate with each other and their communities by using sign language. As this is a natural method for people to interact with computers, many academics are researching it in order to make it simpler and more natural without the use of additional gadgets. So, the main objective of gesture recognition research is to develop systems that can recognise certain human gestures and use them, for instance, to communicate information. Quick, extremely accurate hand detection and real-time gesture identification must be possible with vision-based hand gesture interfaces. Learning sign movements is the first step in creating words and sentences for computer-assisted sign language interpretation. Both static and dynamic sign actions are available. Both types of gesture recognition are crucial to human culture, even if static gesture recognition is easier than dynamic gesture recognition. This paper outlines the processes needed to recognise sign language. The method of gathering data, as well as its preprocessing, transformation, feature extraction, categorization, and outcomes, are all examined. For this area of study, there are also some suggested directions for further research. The visual-manual modality used by sign languages is a collection of established languages. When a hand motion is inputted, the system instantly displays the appropriate recognized character on the monitor screen. In the following, research projects that led to a system that uses convolutional neural networks to identify handwriting on the basis of depth pictures the camera collects.**

Keywords*— Convolutional Neural Network, Hybrid classification, Feature extraction

I. INTRODUCTION

People who are deaf or hard of hearing communicate using sign language, a visual-gestural language. Messages are sent through a variety of body parts, hand motions, and three-dimensional places. It has a unique vocabulary and grammar that is completely distinct from spoken and written languages. Spoken languages employ oratory to create sounds that are connected to particular words and grammatical combinations in order to transmit significant information. Following that, the auditory faculties analyse the oratory components as necessary. Rather than being a spoken language, sign language is a visual one. The grammar of sign language is complex, much like the grammar of spoken language, which uses rules to produce complete messages. A sign language recognition system consists of an accurate method for translating sign language into text or voice. The alphabet flow is recognised, and sign language words and phrases are translated using computerised image processing and a wide range of classification techniques. The use of hand gestures, head motions, and body positions to convey information is common in sign language. There are four key parts that make up a gesture recognition system: gesture simulation, gesture analysis, gesture recognition, and gesture-based application systems. Although it has been studied for many years, sign language recognition might be a game-changer for helping the deaf-mute population. However, each analysis includes flaws and is only a square. It is still not suitable for commercial usage. Several language recognition studies have been successful in the past, but they require a high price to be commercialised.

For creating language recognition systems that will be utilised commercially, researchers are currently receiving more attention. In many different ways, researchers conduct their studies. Beginning with information gathering methods. The cost of a competent device influences the information collecting process, yet a low-cost methodology is necessary for the language recognition system to be commercially successful. Researchers use a variety of techniques for creating language recognition. Every methodology has its own advantages over other approaches, and academics continue to use many approaches while creating their own language recognition. Each methodology also has drawbacks when compared to other approaches. The purpose of this work is to examine several language recognition techniques and identify the most straightforward way used by researchers. As a result, other researchers will gather more information about the methods employed and create superior language application systems in the future. Disabled People employ non-verbal communication, such as these sign language movements, to communicate their thoughts and feelings to other regular people. Yet, skilled sign language interpreters are required for medical and legal consultations as well as training and educational sessions since it is so difficult for the typical person to grasp what they are saying. The need for these services has increased during the previous several years. Several services, including video remote human interpretation, are being developed with the use of high-speed Internet connections. These services offer a fundamental sign language interpreting service that is beneficial, but has significant drawbacks. This project's primary goals are to advance automatic sign language translation and text or speech recognition. We concentrate on static hand motions in our effort for sign language.

The intricacy of the issue prevents the creation of an entirely automatic Sign Language transcriptor system, despite the fact that research on machine-based Sign language recognition has long been ongoing. Due to the size of the problem, it is impossible to find a simple solution (number of motions and variations, movement, facial expressions, and contextual significance). This is one of the explanations for why prior studies concentrated on small datasets of movements and users. Over hundred different concepts are represented by the sign language's many symbols. Depending on the circumstance and the performer's intended message, these signals may be changed in a number of ways, including by pointing, moving the arm, or changing the performer's facial expression. For an example of how the meaning of a

sign could vary, consider the term "not yet". The tongue must touch the lower lip while the head must be tilted to one side in order to sign a phrase. It will be difficult to interpret the sign until one of them is finished.

II. Literature Survey

[1] Ambekar, A. G., Nikam, A. S. (2016), they suggested that sign language interpretation utilizing image-based hand gesture recognition algorithms can be achieved by matching hand template examples while considering contour curve forms into consideration. It is simple to identify hand gestures using contour analysis, irrespective of shape or size, by considering vector values into consideration.

[2] Shrenika (2020), here, the system is implemented utilising image-processing methods. For those who have no way to use gloves, sensors, or other sophisticated equipment, this device is accessible. Snap a picture with a camera first. Finally, for additional processing, convert it to a grayscale image. The image's symbol was located using an edge detection method. The sign alphabet is shown as the final stage.

[3] Talukder (2020), this work seeks to advance the wellbeing of persons who are physically restricted. The proposed method instantly converts sign language into sentences, then speech. Because to the system's ability to collect more than 30 frames per second, a high-efficiency signature may produce assertions instantly. Also, it can help persons who are deaf or mute with their everyday communication needs.

[4] Martinez-Guevara (2019), if it is practicable, phonetic units in the Mexican language will be categorised in the paper according to the motions used to pronounce it. The identification of a categorization that hasn't been found in the field of linguistics using this process may be achievable. Due to the high information processing and significant data quantities, we predict that this procedure may become monetarily and computationally costly. Additionally, it serves as a starting point for the development of tools that will make computational discourse analysis of sign language easier.

[5] Hoque (2018), in this work, we explore the possibility of categorising phonetic units in the Mexican language in accordance with the movements used during pronunciation. By doing this, we might be able to name a hitherto unnamed category in linguistics. Due to the high information processing and the enormous volumes of data, we assume that this process may become time-consuming and computationally expensive. Moreover, it acts as a foundation for the development of tools that will enable the computational analysis of sign language dialogue.

[6] Naglot D., & Kulkarni M. (2016), using the recently unveiled Leap Motion Controller, this article describes a technique for American Sign Language detection. Hands, fingers, bones, and items that resemble fingers may be tracked and recognised using a 3D, non-contact motion sensor known as a Leap Motion Controller. Exact motion and location reports are available from it.

[7] Abiyev R., Idoko J. B., & Arslan M. (2020), the cross validation strategy is applied to train the translator. Nevertheless, only the results of the test with the best conditions are presented in this study, even though several experiments were carried out. The RMSE was 0.0234 and the RMSE of the model was averaged at 92.21%. Studies demonstrate that the suggested sign language translation method is useful for categorising texts that contain ASL fingerspelling. Moreover, experiments revealed that the sign labelling algorithm was able to identify and distinguish between a variety of signs quickly and accurately.

[8] Htet S. M., Aye B. & Hein M. M. (2020), in order to categorise datasets of photographs in Myanmar Sign Language, this study recommends using convolutional neural networks (CNN). It also addresses the skin color enhancement approach for skin detection and the Viola Jones algorithm for face detection. Nonetheless, the Viola Jones algorithm can only recognise frontal images. The suggested approach uses machine learning to be able to identify motion signs in Myanmar Sign Language.

III. Proposed Methodology

A. About the Dataset

The dataset consists of different images of hands for the recognition of the sign input provided. For example, image of no, yes, okay. Data set collecting comes first in the process. The Kinect camera's full-frame (RGB) picture and accompanying depth map are obtained. An infrared projector and a monochrome CMOS sensor are included within the 3D depth sensor, helping to produce a simulated 3D image throughout its field of vision. By emitting invisible near-infrared rays and measuring the time it takes for those rays to travel after reflecting off an item in its area of view, it can determine how far away each point in its field of view is from the center.

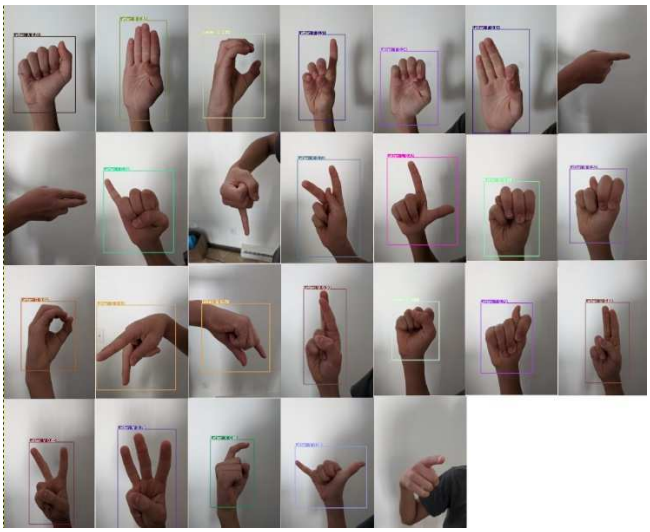

Fig. 1. Dataset

B. Convolutional Neural Network

Convolutional neural network is the term used to describe a multilayered feedforward neural network (CNN). It is one of the methods for classifying photos that is most frequently applied. It includes a number of pooling layers, activation functions, and convolution layers. It ends at the layer that is entirely linked. The predicted label is produced by the CNN using an image vector representation as its input. The number of neurons in the completely connected layer is proportional to the number of classes. The efficiency of the algorithm is also affected by a loss function.

- Convolution is used to draw out certain details from an image's input. By employing tiny squares of input data to train an image feature model, it retains the spatial connection between pixels. In most cases, relu comes next.
- The relu operation replaces each negative pixel value in the feature map with a value of zero, element by element. Its goal is to non-linearize a convolution network.
- Although each feature map's dimensionality is decreased by pooling, which is also known as downsampling, significant data is kept.
- The fully linked layer is a multi-layer perceptron, where the output layer has the softmax function. It divides the input image into several groups depending on characteristics from earlier layers and training data.

A CNN model is produced using the intersection of these layers. A fully linked layer is the last layer.

Assume our convolutional layer comes before a layer of N×N square neurons. Our convolutional layer output, if we employ a m×m filter ω, will have the following size: (N−m+1)×(N−m+1). The contributions from the preceding layer cells must be added together (weighted by the filter components) in order to determine the pre-nonlinearity input to a particular unit in our layer:

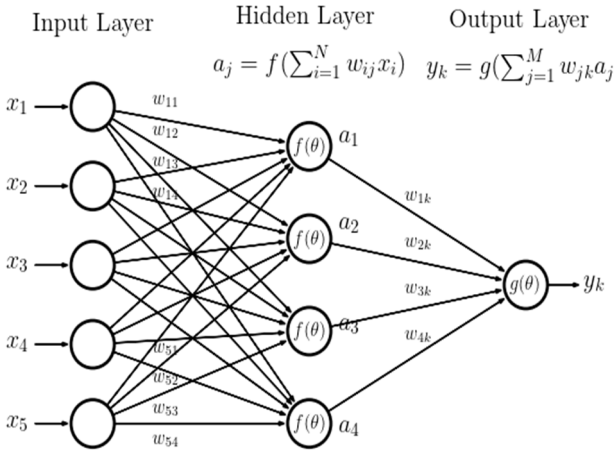

Fig. 2. Layers in Convolutional Neural Network

C. K – Nearest Neighbor

A supervised learning approach that classifies objects in feature space is the k closest neighbor algorithm. Each row of data in the sample is assigned using the KNN technique using the nearest-neighbor algorithm to one of the training groups. Using a variety of training photos, the category is a collection of traits that were learned throughout the training phase. The main goal of classification is to select the best reference characteristic feature vector from those that best matches the new feature vector. In order to obtain the KNN in a dimensional space, it uses feature vectors produced during the training phase. An overwhelming majority of the vector's neighbors decide on its categorization. A selection of objects for which the right categorization has been established are used to determine neighbors.

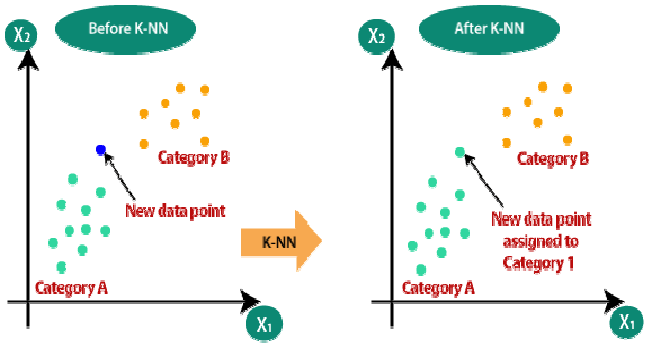

Fig. 3. K-Nearest Neighbors

IV. Proposed Workflow

A. Pre-processing

To extract the region of interest from the training images, a preliminary processing step is carried out (ROI). If just hand motions are taken into account, the ROI may be the hands; however, if facial emotions are also taken into account, the ROI may be both the face and the hands. Techniques used for pre-processing include morphological filtering, segmentation, picture scaling, and filtering. We can use any of the widely used methods for image filtration and enhancement. Otsu's thresholding, background removal, segmentation based on skin colour and motion are a few techniques utilised for segmentation. The test photos or videos are also pre-processed to enable isolation of the area of interest during the testing stage.

- In Image Enhancement photographs were enhanced in terms of colour, brightness, sharpness, and contrast. As an illustration, the brightness and contrast were adjusted to allow for the differentiation of fingers even in extremely dark images.
- Edge enhancement sharpens edges, This is accomplished by raising the contrast in an area of the image that is locally identified as an edge. The line between the hand and fingers and the background becomes considerably more obvious as a result. This could make it easier for the neural network to recognise the hand and its limits.
- In Image Whitening the singular value decomposition of a matrix is used in the ZCA, also known as picture whitening, approach. This method decorrelates the data and eliminates any unnecessary or apparent information. As a result, the neural network may search for more intricate associations and learn about the underlying structure of the patterns it is trained on. The mean is zero, and the covariance matrix of the picture is set to identity.

B. Feature Extraction

In this vision-based technique, the fingers, joint angles, and palms are the features calculated and used to accomplish recognition. Deep Neural Network is used to extract features and reduce dimensionality of the data. As it creates the feature vectors that are utilised as the classifier's inputs, the feature extraction phase is among the most crucial ones in the recognition of sign language. The ability of feature extraction algorithms to accurately identify shapes should be

independent of how the lighting, the orientation of the objects, or their size change in a video or photo. There are several methods that may be used to construct the features, such as Fourier descriptors, orientation histograms, and scale invariant feature transformations. Then, using one of the feature extraction approaches to acquire the feature vector, the classifier is trained using that information.

C. Classification

In order to divide the input signals into distinct classes for sign language identification, a classifier is required. Using the feature vector that was gathered from the training database, the classifier is trained during the training phase. The trained classifier recognises the class of the sign when a test input is supplied, and then plays or displays the pertinent text in line with that determination. As test inputs, you can utilise images or videos. They come into two categories: supervised and unsupervised machine learning methodologies. The supervised machine learning approach may be used to educate a computer to recognise patterns in input data that can then be used to the prediction of future data. In order to determine a function using training data that is labelled, supervised machine learning employs a set of predetermined training data.

To derive conclusions from datasets with unlabelled answers, unsupervised machine learning is required. The classifier does not receive a labelled response, thus no reward or penalty weighting is used to determine which classes the data is meant to be placed in. Some of the most popular classifiers include fuzzy systems, artificial neural networks, multiclass support vector machines and K Nearest Neighbor (KNN). The classifier's effectiveness is measured by its recognition rate.

D. Testing

Around 1,000 images of each sign is used to determine if the preprocessing of the photographs did actually result in a more reliable model. The data was divided in test set and training set in the ratio of sixty and forty percent, we evaluated our hypothesis using a test set made up of both our own image data and images from the original dataset. We find that perhaps the model that are trained on preprocessed images outperforms the model trained on the original images, most likely because the former exhibits less overfitting.

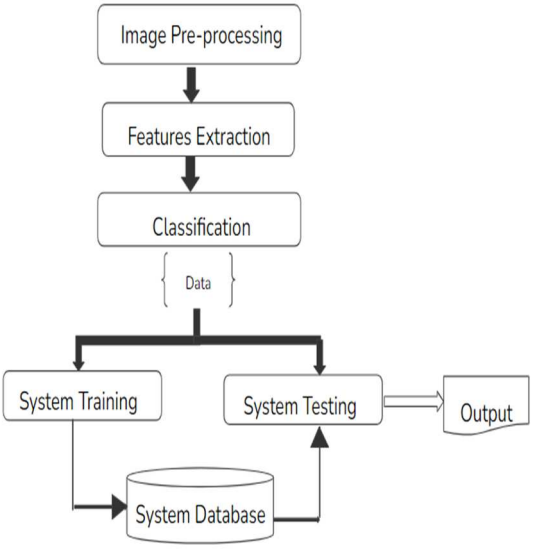

Fig. 4. Architecture Diagram

V. Result and Output

How well our feature extraction technique works. Convolutional neural networks are used in the proposed system for the recognition of sign language, which is implemented on the Microsoft Kinect platform. The system also makes use of GPU acceleration. Convolutional Neural Networks flawlessly automate the crucial process of feature extraction. It is used with the Sign Language dataset, which contains a large number of movements. Before the input is sent to the neural network, the system must pre-process it. The high hand and its rectangular form are resized as the first phase in the pre-processing. Also, the motions of the fingers are taken into account. After pre-processing, four video samples with a resolution of 64x64x32, or 32 frames of size 64x64, are produced. Following extensive pre-processing and training on the massive dataset, it was discovered that 95% of the time the product correctly recognised the meaning of the shown hand gesture. The 5% mistake rate is caused by a number of different factors, including poor lighting, fuzzy images, and issues in capturing images from video streams.

In general, the system performs as predicted. So, it may be claimed that it'll be successful. Yet there is still a lot of room for improvement. The efficacy of this suggested process would be improved while we continuously worked to reduce mistake rates.

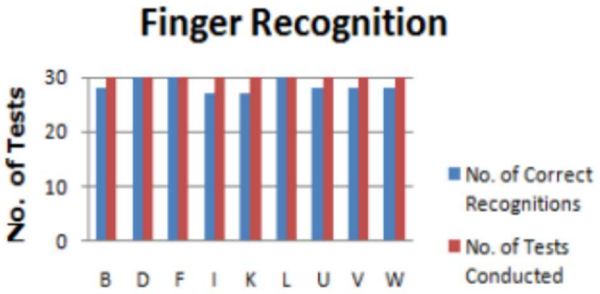

Fig. 5. Accuracy of Sign Language Gestures

TABLE I. Recognition Result Analysis

GESTURES	NO OF TEST CONDUCTED	NO OF CORRECT RECOGNITION
B	30	30
D	30	28
F	30	29
I	30	26
K	30	28
L	30	28
U	30	30
V	30	28
W	30	29

An effective system for recognising hand gestures has been built for this project, along with a method for generating commands based on gestures. With the aid of a webcam, the suggested system is practical and can record gestures in real time. Using spatial or temporal invariance in recognition is the reason CNN was chosen for picture classification. CNN outperformed previously employed methodologies in terms of efficacy. Convolution layer and pooling layer are two crucial aspects of CNN that set it apart from conventional neural networks. The segmentation of the gesture was carried out using geometrically based compactness detection with regard to backdrop references. The segmentation of gestures is based on how compactness varies in relation to particular predefined criteria. The left half of the hand's compactness in terms of the area it occupies in the palm and the radial distance are the two main parameters that the overall system takes into account. A final real-time interpretation of that specific hand motion is formed by combining all of these variables.

When these results are shown on a graph, the number of tests is represented by the Y axis, and the American Sign Language motions are depicted on the X axis in relation to the alphabets that the user is searching for. We couple the columns that are present for each gesture.

- The first column displays the number of times the algorithm accurately identified the gesture's meaning.
- The second column shows how frequently a user has used the system under test.

VI. Conclusion

Consider all the many gesture combinations that a system of this type must be able to comprehend and communicate; decoding sign language is a challenging undertaking. Considering that, it is perhaps best to split this problem into smaller problems, with the approach presented here acting as a potential solution to one of them. As demonstrated in the demonstration, first-person sign language translation systems may be made using convolutional neural networks and just cameras. This will be followed by a review of the solution and research for system upgrades. There are several strategies to make improvements, including the testing of different convolutional neural network architectures, redesigning vision systems, and gathering higher-quality data. To close the communication gap that persons with hearing impairments experience, particularly in the digital age, this is a critical issue to be resolved. In conclusion, we were able to create a workable and valuable system that can comprehend sign language and convert it to the equivalent text.

VII. Future Scope

In order to identify ISL words and phrases, we can develop a model. For this, you will need a system that can spot changes in respect to temporal space. The communication gap can be closed by developing a complete solution that will help the deaf and hard of hearing. Later on, the Raspberry Pi platform may be used to create and implement the recommended system. We can try to identify motion-related clues. Additionally, we may concentrate on transforming the series of motions into text, or words and sentences, and finally becoming audible voice. It would take a server with a lot of Memory and storage capacity to use our initial dataset because of its size. The training, validation, and test sets' file names might be divided, and the Dataset class's photos could be dynamically loaded as a potential solution. We would be able to train the model on more dataset samples if we used such a loading strategy.

References

[1] Shrenika, S., & Madhu Bala, M. (2020). Sign Language Recognition using Template Matching Technique. 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA).

[2] Talukder, D., & Jahara, F. (2020). Real-time Bangla Sign Language Detection with Sentence and Speech Generation. 2020 23rd International Conference on Computer and Information Technology (ICCIT).

[3] Martinez-Guevara, N., Rojano-Caceres, J.-R., & Curiel, A. (2019). Detection of Phonetic Units of the Mexican Sign Language. 2019 International Conference on Inclusive Technologies and Education (CONTIE)

[4] Hoque, O. B., Jubair, M. I., Islam, M. S., Akash, A.-F., & Paulson, A. S. (2018). Real Time Bangladeshi Sign Language Detection using Faster R-CNN. 2018 International Conference on Innovation in Engineering and Technology (ICIET).

[5] Naglot, D., & Kulkarni, M. (2016). Real time sign language recognition using the leap motion controller. 2016 International Conference on Inventive Computation Technologies (ICICT).

[6] Abiyev, R., Idoko, J. B., & Arslan, M. (2020). Reconstruction of Convolutional Neural Network for Sign Language Recognition. 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE).

[7] Htet, S. M., Aye, B., & Hein, M. M. (2020). Myanmar Sign Language Classification using Deep Learning. 2020 International Conference on Advanced Information Technologies (ICAIT).

[8] Konwar, A. S., Borah, B. S., & Tuithung, C. T. (2014). An American Sign Language detection system using HSV color model and edge detection. 2014 International Conference on Communication and Signal Processing.

[9] Hossein, M. J., & Sabbir Ejaz, M. (2020). Recognition of Bengali Sign Language using Novel Deep Convolutional Neural Network. 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI).

[10] Bhadra, R., & Kar, S. (2021). Sign Language Detection from Hand Gesture Images using Deep Multi-layered Convolution Neural