Sign Language To Speech Conversion Using Deep Learning

Atharva Shinde[1], Anushri Shivale[2], Siddhesh Phapale[3], Assistant Prof. Renuka Kajale[4]

[1,2,3]Students, CS Department, NMIET, Talegaon, Pune, India [4]Assistant Professor, CS Department, NMIET, Talegaon, Pune, India

***Abstract* :** Through communication, people can engage and share thoughts and feelings. There are several obstacles in the way of the deaf community's social interactions. The individuals use sign language to communicate with one other. In order to communicate with regular people, a technology can convert sign languages into a form that is understandable. The goal of this project is to create a real-time text-to-Indian Sign Language (ISL) translation system. Most of the work is done by hand. In this paper, we present a deep learning technique for classifying signs using a convolutional neural network. Using the numerical signs and the Python-based Keras convolutional neural network implementation, we first build a classifier model. Phase two involved using a second real-time system that located the Region of Interest in the frame that displays the bounding box using skin segmentation. The segmented region is fed into the classifier model in order to forecast the sign. For the identical subject, the system's accuracy rate is 99.56%; in low light, it is 97.26%. The classifier was seen to be becoming better with different background and angle of image capture. Our approach focuses on the RGB camera system.

Keywords— Deep learning, convolutional neural networks, regions of interest, and real-time systems.

I. INTRODUCTION

When it comes to physically disabled people, both the deaf and dump communities employ various sign languages. In the world, people to offer communication utilize many languages. American Sign Language, Chinese Sign Language, Indian Sign Language, and others are among the several sign languages. In each instance, the symbols alter depending on whether motion, single-handed, or double-handed representations are present. In some circumstances, dynamic symbols are utilized for words like "hello," "Hai," etc. instead of static symbols to represent letters. These communities' ability to communicate with one another will be enabled through a real-time system. It can be converted to any language after being transformed using the Computer Vison method. To create an accurate and efficient system, numerous studies have been conducted in this area. The researchers' earlier approach utilized a handcrafted feature, but it was constrained and used under particular circumstances. The majority of works rely on feature extraction based on HOG, SIFT, LBP, etc., as well as pattern recognition. However, most of the time a system using just one feature is insufficient; hence the hybrid technique was developed to address this issue. But in a real-time system, we need to solve problems more quickly. Nowadays, we use parallel implementation to increase the processing speed of our computers. Our system uses a single core to solve issues the majority of the time. Parallel computing can be used to solve problems using the GPU system, which has more cores than a CPU system. We can model a self-learning system for our needs using the deep learning methodology. One of the most popular deep learning systems that can handle any computer vision issue is the convolutional neural network. For the real-time implementation of our technique, we used a region of interest convolutional neural network.

1.1 Project Domain Description

Because hand signs are used by the dumb to communicate, normal people have trouble understanding what they are trying to say. As a result, technologies that can distinguish between various indications and alert the general public are required.

1.2 Application

The primary application is to provide a means of communication for individuals with hearing impairments. This technology enables them to interact with others who may not understand sign language by converting their gestures into spoken words. In educational settings, this technology can facilitate communication between students who use sign language and teachers or peers who do not. It can also be integrated into e-learning platforms to provide accessible content for online courses. Businesses can implement this technology in customer service environments to provide support to customers who communicate through sign language, enhancing inclusivity and accessibility.

II. LITERATURE SURVEY

One of the most popular trends in technology today is computer vision, which is used in many AI-based systems including robots, cars, markets, etc. Concerns about object detection and image classification are more influenced by the system. This technique can be used to implant the sign language system. In the earlier systems, numerous more techniques were employed. [1]. made use of literature as a basis for the ISL Recognition system. A glove is used for color segmentation, and Principal Component Analysis, or PCA, is used for recognition. Real-time data frames are used as input for recognition every 20th frame. This approach has problems with the sign's mobility and overlapping. PCA and the fingertip algorithm are both utilized for recognition. Recent studies have concentrated on static indicators of ISL [2]. from photo or video sequences that were captured using data glove or colored glove under controlled circumstances such a single background and specialized gear. In the system, the light and position are more significant. To work under these circumstances, the signer needs to be knowledgeable of the system. Preprocessing Otsu's thresholding can be done in a variety of ways, including by considering skin tone, motion-based segmentation, and backdrop subtraction [3–5].Scale-invariant feature transform, Fourier descriptors, and wavelet decomposition are used in the feature extraction phase. K Nearest Neighbor (KNN), Hidden Markov Models (HMM), Multiclass Support Vector Machines (SVM)[6], Fuzzy systems, Artificial Neural Networks (ANN), and many other classifiers are used to categorize signs. implemented an edge detection approach for hand gesture identification in another study [7]. Edge detection and sorting characteristics in the database are used to retrieve the frame features. With the freshly constructed database, predict the gesture by applying template matching. In this instance, templates are matched using the least distance. The system is capable of identifying both static symbols and dynamic gestures. Using a fuzzy membership function and a fuzzy [8]. based approach, the system extracts the spatial aspects of signs. The Nearest Neighbor classifier is paired with a suitable symbolic similarity measure. Reheja [9]. et al. developed an Indian sign language gesture detection system using the Microsoft Kinect sensor device. They used RGB and Depth Kinect pictures for their investigations [2].The study indicates that using RGB-D images increases the accuracy of the system. The HU-Moments, which are moments that are angle, position, and shape invariant, are extracted by the model and fed to the SVM classifier as features.Indian sign language has an android app-based system designed by Pranali Loke[10]. et.al.Images are collected by the Android system and sent to the server.The server system sends these photos to the Matlab application, where the system is trained using a neural network and features are extracted using the Sobel operator.The system analyzes the photos using pattern recognition and classification to produce text as the result. Beena M.V. developed a technique to identify American Sign Language (ASL) from depth photos taken by the Kinect sensor.et al.A total of 1000 photographs of each numerical sign were used to train the system [11].The approach produced a 99.46% accuracy for the depth pictures after extracting features from the block-processed images and training an artificial neural network (ANN).On the, the system has been taught for quicker execution.Convolutional Neural Network [2017-2] (CNN) with softmax classification 1 is applied as a continuation of the work for 33 static Kinect depth picture symbols.The way it's implemented shows that the custom feature become insufficient for classification purposes as the number of classes rises. The CNN structure will perform better in terms of accuracy compared to other conventional methods because it can learn from the provided training data.

III. METHODOLOGY

The method used by the system is vision-based. Since all indications can be read with just the hands, the need for artificial devices to facilitate interaction is eliminated.

3.1 Dataset Generation

We looked for pre-made datasets for the project, but we were unable to locate any that met our needs for raw image format. The RGB value datasets were the only ones we could locate. We thus made the decision to produce our own data set. The following are the steps we used to construct our data set. We created our dataset using the Open Computer Vision (OpenCV) framework. First, for training purposes, we took about 800 pictures of each symbol in ISL (Indian Sign Language), and for testing, we took about 200 pictures of each symbol. Initially, we record every frame that our machine's webcam produces. We designate a Region Of Interest (ROI) in each frame, which is represented in the image below by a blue bounded square:

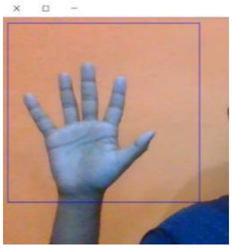


Fig. 1 ROI Example Image

Then, we apply Gaussian Blur Filter to our image which helps us extract various features of our image. After applying Gaussian Blur, the image appears as follows:

Fig. 2 Gaussian Filter

Then the algorithm is applied.

IV. ALGORITHM

In the realm of computer vision, a popular deep learning architecture is the convolutional neural network (CNN). Computers can now sense and evaluate visual data, such as photos, thanks to a field of artificial intelligence called computer vision. Machine learning, including artificial neural networks, has shown remarkable performance in various domains. Many datasets, such as text, audio, and image datasets, are used with neural networks. Various neural network topologies are used for different applications. For example, recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, are employed in natural language processing jobs to anticipate word order. Convolutional neural networks are also frequently used for image categorization applications.

In contrast to conventional neural networks, CNN's layers include neurons arranged in three dimensions: width, height, and depth. Rather than being entirely coupled to every other neuron in the layer, a layer's neurons will only be connected to a tiny portion of the layer (window size) preceding it. Furthermore, as the entire image will eventually be reduced to a single vector of class scores at the end of the CNN architecture, the final output layer would have dimensions (number of classes).

4.1 Convolutional Layer:

We use a modest window size in the convolution layer that extends to the input matrix's depth, usually measuring 5 by 5. The layer is made up of window-sized learnable filters. In each iteration, we compute the dot product of the input values at a given place and slid the window by a stride size, usually 1. As we go, a 2-Dimensional activation matrix that displays the matrix's reaction at each spatial place will be produced. In other words, the network will pick up filters that turn on when it detects certain kinds of visual features, such a splotch of a particular color or an edge with a particular orientation.

4.2 Pooling Layer:

In order to lower the size of the activation matrix and, eventually, the learnable parameters, we employ a pooling layer. Two categories of pooling exist:

a. Max Pooling: This technique uses a window size, such as a 2*2 window, and only takes the highest four values. We'll close this window and carry on until we eventually get an activation matrix that is half the size it was originally.

c. Average Pooling: This method makes use of every Value within a window.

4.3 Fully Connected Layer:

In a fully connected region, all inputs will be connected to neurons, however in a convolution layer, neurons are only connected to a small region.

4.4 Final Output Layer:

We will connect the information from the completely connected layer to the last layer of neurons, whose count equals the total number of classes, so that it can forecast the likelihood that each image will belong to a different class.

4.5 Data Flow:

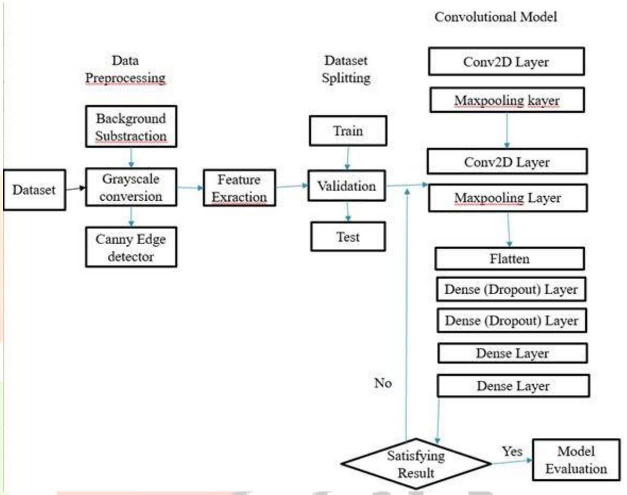


Fig. 3 Flowchart

V. ADVANTAGES

1) Improved Communication: Sign language to speech conversion technology allows those who are deaf or hard of hearing to communicate successfully with those who do not understand sign language, increasing their capacity to connect and exchange ideas with the larger community.

2) Real-time Translation: The creation of a real-time system for sign language translation enables instant communication between sign language users and non-signers, allowing for natural and fluid interactions with little delays or impediments.

3) Accessibility: Sign language to voice conversion technology makes information more accessible to those with hearing impairments by enabling alternate modes of

communication, reducing their reliance on interpreters and textual communication.

4) Adaptability: The system's ability to enhance classification accuracy despite changing backgrounds and image collection angles illustrates its adaptability to multiple environmental conditions, making it appropriate for a wide range of real-world applications.

VI. FUTURE SCOPE

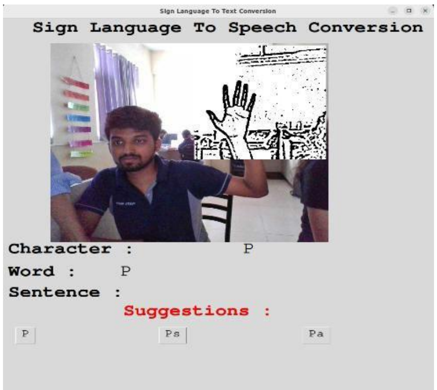
1) Expansion to Other Sign Languages: Although the project concentrates on translating Indian Sign Language (ISL) into text, there is potential to expand the system to cover other sign languages used throughout the world. This would entail gathering data for various sign languages, training models, and tailoring the system to accommodate linguistic and cultural differences.

2) Gesture Recognition and Translation: In addition to identifying individual signals, future studies may concentrate on detecting and translating complex gestures, facial expressions, and non-manual markers used to transmit meaning in sign languages. To capture the intricacies of sign language communication, advanced deep learning algorithms and multimodal integration are required.

3) Mobile and Wearable Applications: Creating mobile apps or wearable devices that offer on-the-go access to sign language translation services can improve accessibility and convenience for those with hearing impairments. This would entail optimizing the system for mobile platforms and creating user-friendly interfaces customized to the needs of mobile users.

VII. RESULT

On the other hand, real-time performance did not meet the expected standards, despite the rather strong training accuracy (around 99%) obtained when training the picture dataset without any augmentation. Because hand gestures and signals were not perfectly centered and vertically aligned in real time, it was not predicted with accuracy most of the time. We enhanced our dataset to overcome this restriction and raise our model's efficiency. The training accuracy dropped to 89% as a result, but the real-time forecasts remained mainly accurate. In offline testing, 92.7% accuracy was shown utilizing about 9000 augmented photos. Furthermore, during both training and real-time application, the accuracy tends to grow with the number of parameters in the model.



Sign Language To Speech Conversion

Character : P

Word : P

Sentence :

Suggestions :

Fig. 4 GUI

VIII. CONCLUSION

The real-time system has been built for numeral signals from 0-9. This is the first step towards the recognition of Indian Sign Language. The first step towards Indian Sign Language being recognized is this. RGB's 3000 static symbols. For testing, the system used 100 photos for each symbol. The model was developed through the effective use of a region-based convolutional neural network in a deep learning system. For the same subject, the system achieved an accuracy of 99.56% during testing, but in low light, the accuracy dropped to 97.26%. Add more symbols from the alphabets of the Indian sign language's static symbols in the future, including the double hand notation. The dataset must be expanded in order to address the low light issues.

REFERENCES

[1] Divya Deora, Nikesh Bajaj, Indian Sign Language Recognition, 2012 1st International Conference on Emerging Technology Trends in Electronics, Communication and Networking, IEEE 2012-978-1-4673-1627-9/12.

[2] Anuja V. Nair, Bindu V., A Review on Indian Sign Language Recognition, International Journal of Computer Applications (0975 – 8887) July 2013, Volume 73– No.22

[3] Jorge Badenas, Josee Miguel Sanchiz, Filiberto Pla, Motion-based Segmentation and Region Tracking in Image Sequences, Pattern recognition 2001, 34, pp. 661-670.

[4] Ping-Sung Liao, Tse-Sheng Chen, Pau-Choo Chung, 2001, A Fast Algorithm for Multilevel Thresholding, Journal of Information Science and Engineering 17, pp. 713-727.

[5] Dr. Alan M McIvor, Background subtraction techniques, Image and Vision Computing, Newz Zealand 2000 (IVCNZ00).

[6] Aseema Sultana, T. Rajapushpa, Vision Based Gesture Recognition for Alphabetical Hand gestures Using the SVM Classifier, International Journal of Computer Science and Engineering Technology, Volume 3, No. 7, 2012.

[7] Purva A. Nanivadekar, Dr. Vaishali Kulkarni, Indian Sign Language Recognition: Database Creation, Hand Tracking and Segmentation, International Conference on Circuits, Systems, Communication and Information Technology Applications, IEEE 2014,978-1- 4799-2494-3/14.

[8] Nagendraswamy H S, Chethana Kumara B M, Lekha Chinmayi R, Indian Sign Language Recognition: An Approach Based on Fuzzy-Symbolic Data, 2016 Intl. Conference on Advances in Computing,

Communications and Informatics (ICACCI), Sept. 21-24, 2016, 978-1-5090-2029-4/16.

[9] J. L. Raheja , A. Mishra, A. Chaudhary, Indian Sign Language Recognition Using SVM, Pattern Recognition and Image Analysis, 2016, Vol. 26, No. 2, pp. 434–441.

[10] Pranali Loke, Juilee Paranjpe, Sayli Bhabal, Ketan Kanere, Indian Sign Language Converter System Using An Android App. ,International Conference on Electronics, Communication and Aerospace Technology, 2017 IEEE ,978- 1-5090-5686- 6/17.

[11] M.V. Beena and M.N. Agnisarman Namboodiri, ASL Numerals Recognition from Depth Maps Using Artificial Neural Networks, Middle-East Journal of Scientific Research 25 (7): 1407-1413, 2017,ISSN 1990-9233.

[12] Beena M.V., Dr. M.N. Agnisarman Namboodiri, Automatic Sign Language Finger Spelling Using Convolution Neural Network: Analysis, International Journal of Pure and Applied Mathematics, Volume 117 No. 20 2017, 9-15.