INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1560)
Available online at: https://www.ijariit.com

Sign Language to Text and Speech Conversion

Bikash K. Yadav
bikashyadav0055@gmail.com
Sinhgad College of Engineering, Pune, Maharashtra

Dheeraj Jadhav
dhirajadhav.56@gmail.com
Sinhgad College of Engineering, Pune, Maharashtra

Hasan Bohra
hasanaarif22@gmail.com
Sinhgad College of Engineering, Pune, Maharashtra

Rahul Jain
rahuljain199824@gmail.com
Sinhgad College of Engineering, Pune, Maharashtra

ABSTRACT

Sign language is one of the oldest and most natural forms of language for communication. Since most people do not know sign language and interpreters are very difficult to come by, we have come up with a real-time method using Convolution Neural Network (CNN) for fingerspelling based American Sign Language (ASL). In our method, the hand is first passed through a filter and after the filter has applied the hand is passed through a classifier that predicts the class of the hand gestures.

Keywords— *CNN, ASL*

1. INTRODUCTION

American Sign Language (ASL) could be a total, normal dialect with etymological highlights comparable to talked languages and linguistic use unmistakable from English. Hand and confront developments are utilized to precise ASL. It is the overwhelming language of numerous deaf and hard-of-hearing people in North America, as well as numerous hearing people. There's no formal or broadly recognized sort of sign dialect. Distinctive flag dialects hypothesize around particular themes. For illustration, British Sign Language (BSL) could be a totally distinctive dialect than American Sign Language (ASL), and people who are commonplace with ASL will battle to get a handle on BSL. ASL was not made by a single individual or a bunch of individuals. The precise beginnings of ASL are obscure, in spite of the fact that a few conjectures that it emerged over 200 a long time back from the blending of neighborhood sign dialects and French Sign Language (LSF, or Langue des Signes Française). Today's ASL combines parts of LSF and the primary nearby sign languages; all through time, these have combined and advanced into a wealthy, complex, and modern language. ASL progressed and LSF progressed are two particular languages. Whereas they still have many signs in common, they can now not be recognized by each other's clients. The American Sign Language (ASL) could be a totally special and autonomous language from English. It has its claim rules for elocution, word creation, and word arrangement, as well as all of the basic highlights of the language. Languages vary in how they flag particular capacities, such as inquiring an address instead of conveying an explanation. ASL incorporates fingerspelling, which is utilized to sign out English words. Each letter within the fingerspelled letter set compares to a distinctive hand shape. Fingerspelling is habitually utilized to mean appropriate names or the English term for something. The database contains one thousand special gesture graphics, taking into account all of the sign language alphabets and phrases. The suggested system aims to comprehend some very basic signal language elements and translate them to text and voice. American Sign Language is a language that can be seen. Along with signing, the ideas use vision to process language data. The shape, positioning, and motion of hands, as well as facial emotions and frame motions, all play an important role in communicating information. Sign language isn't a typical language - it's used by every single person in the United States. It has its own signal 6 languages, and places have linguistic quirks, just as many languages are spoken anywhere in the world. The detection rate of the ASL language, when compared to grammatical accuracy, is 90%.

2. LITERATURE SURVEY

Paper [1] demonstrates the classification of motion is done using Hidden Markov Models (HMM). The dynamic features of gestures are addressed in this approach. The skin-color blobs corresponding to the hand are tracked into a body–facial space centered on the user's face to extract gestures from a succession of video images. The purpose is to distinguish between two types of gestures: deictic and symbolic movements. A rapid lookup indexing table is used to filter the image. Skin color pixels are grouped into blobs after filtering. Blobs are statistical objects used to determine homogenous areas based on the location (x, y) and colorimetry (Y, U, V) of skin color pixels.

Authors of [2] have used Naïve Bayes Classifier, which is an effective and fast approach for recognizing static hand gestures. It works by identifying distinct movements based on

geometric invariants extracted from visual data after segmentation. As a result, unlike many other recognition algorithms, this one is not based on skin color. With a static background, the gestures are retrieved from each frame of the movie. The initial stage is to segment, label, and extract geometric invariants from the objects of interest. To produce adequate data for a locally weighted Naive Bayes classifier, the next step is to classify gestures using a K closest neighbor algorithm augmented with a distance weighting algorithm (KNNDW).

Authors of [3], extract the hand from a picture, create a skin model and then apply a binary threshold to the entire image. They calibrate the threshold image around the primary axis in order to center the image around it. They used this image to train and predict the outputs of a convolutional neural network model. They trained their model on seven hand gestures, and it produces an accuracy of roughly 95% for those movements when used with their model.

Authors [4] propose a method for merging multiple layers' features in CNN models. Furthermore, features from several layers are extracted using previously taught CNN models with training photos. CIFAR-10, NORB, and SVHN image classification benchmark data sets are used to assess the proposed fusion method. The suggested strategy improves the reported performances of the existing models by 0.38 percent, 3.22 percent, and 0.13 percent, respectively, in all three scenarios.

Authors [5] presented the idea of a translation with skin color tone to determine the ASL utilizing machine learning techniques. They created a skin color segmentation that shows the color and assigns a tune to it for further detection. They picked YCbCr color spacing because it's widely utilized in video template code and produces an effective color tone for human skin. They've also used the CbCr plane to spread the color of the skin tone.

The authors [6] have created a system that works in a continuous manner, offering a sequence of sign language gestures to create an automated training set and supplying the spots sign from the training set. With the supervision of noisy texts, they proposed a system that supervises the sentence and figures out the compound sign gesture associated with it using instance learning as a density matrix technique. The collection that was originally designed to demonstrate a continuous data stream of words is now being used as a training set for recognizing gesture posture.

Paper [7] demonstrates the use of a new and standardized type of communication system that primarily targets deaf individuals. The system included the following two scenarios: first, Spanish talks to Spanish sign translation, which required the use of a translator to break apart the words and turn them into a stream of signs that, when combined, create a phrase and are also displayed in avatar form. The second scenario is that the sign language generated from a voice is sent to the generator, which translates the sign language into Spanish words that make sense.

Authors [8] aim to make communication easier by converting ISL signs into voice, which mostly consist of dynamic and static signs. A glove sensor combined with a flex sensor aids in the design of hand orientation and subsequent actions. Wireless transmission is used, and the result is further bits of speech. They investigated LSTM networks that had developed dependencies over time in this project.

Authors [9] create a user-friendly system that is beneficial to those with hearing impairments and who, in general, rely on a very basic and effective method: sign language. The technology can convert sign language to voice, as well as voice to sign language. For sign language conversion, a motion capture system is employed, and for voice conversion, a speech recognition system is employed. It captures the signs and dictates them as writing on the screen. It also records the user's voice and shows the sign language meaning as a moving image or video on the screen.

Authors [10] have suggested a technique for detecting static hand signs of alphabets in American Sign Language automatically (ASL). To do so, they integrated the principles of AdaBoost and Haar-like classifiers. To improve the system's accuracy, they have used a large database for the training procedure, which yielded outstanding results. A data set of 28000 samples of hand sign images, 1000 images for each hand sign of Positive training images in varying scales, lighting, and a data set of 11100 samples of Negative images were used to implement and train the translator. The Logitech Webcam was used to capture all of the positive photographs, and the frames were set to the VGA standard 640480 resolution.

3. METHODOLOGY

The system is based on the concept of vision. All of the signs are made with the hands, which eliminates the need for any artificial equipment for interaction.

3.1 Flowchart

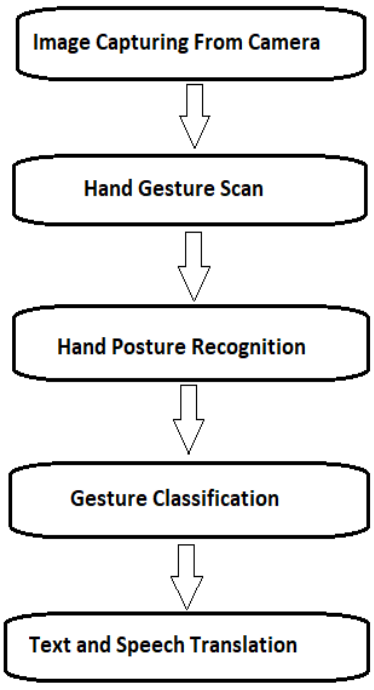

Fig. 1: Flow chart of the project

The flow chart depicts the stages involved in achieving the project's goals.

3.1.1 Image Capturing From Camera: The web camera is used to capture the motions. The full signing duration is captured using this OpenCV video stream. Frames from the stream are retrieved and converted to grayscale images.

3.1.2 Hand Gesture Scan: Hand motions are scanned in the collected images. This is a step in the preprocessing process that occurs before the image is fed into the model for prediction. The passages including gestures have been amplified. This multiplies the likelihood of a correct forecast by a factor of ten.

3.1.3 Hand Posture Recognition: The Keras CNN model is fed the preprocessed pictures. The projected label is generated by the model that has already been trained. All of the gesture labels have a probability associated with them. The anticipated label is determined by the label with the highest likelihood.

3.1.4 Gesture Classification: To predict the final symbol of the user, a variety of symbols that provide comparable consequences gets detected. The classifiers designed specifically for those sets to classify amongst them are used.

3.1.5 Text and Speech Translation: The model translates known gestures into words. The GTTS library is used to convert the recognized words into the appropriate speech. The text-to-speech output is a simple workaround, but it's a useful feature because it simulates a real-life dialogue.

3.2. Convolution Neural Network

CNNs are a type of neural network that is particularly beneficial in the field of computer vision. They were inspired by the actual perception of vision that takes place in our brain's visual cortex. They employ a filter/kernel to scan over the whole pixel values of the image and do computations by assigning appropriate weights to enable feature detection.

3.3 The CNN Model

The resolution of the input image used in this project is 128x128 pixels. It is first processed using 32 filter weights in the first convolutional layer (3x3 pixels each). This creates a 126X126 pixel image for each of the Filter-weights. The images are down sampled using 2x2 max pooling, which means we keep the highest value in the array's 2x2 square. As a result, our image has been reduced to 63x63 pixels. These 63 x 63 pixels from the first pooling layer's output are now used as an input to the second convolutional layer. The second convolutional layer uses 32 filter weights to process it (3x3 pixels each).

As a result, you'll get a 60 x 60 pixel image. The images are then down sampled again using a max pool of 2x2 and reduced to a resolution of 30 x 30 pixels. These images are now fed into a 128-neuron fully connected layer, and the output of the second convolutional layer is reshaped into a 30x30x32 =28800-value array. This layer receives a 28800-value array as input. The 2nd Densely Connected Layer receives the output of these layers. To avoid overfitting, we use a dropout layer with a value of 0.5. The output of the 1st Thickly Associated Layer is presently utilized as an input to a 96-neuron completely associated layer. The output of the moment thickly associated layer is passed into the ultimate layer, which is able to have the same number of neurons as the number of classes we're categorizing. In each of the layers, we used ReLu (Rectified Linear Unit) (convolutional as well as fully connected neurons). For each input pixel, ReLu calculates max(x,0). This gives the formula additional nonlinearity and makes it easier to learn more intricate features. It aids in the elimination of the vanishing gradient problem as well as the acceleration of training by lowering calculation time.

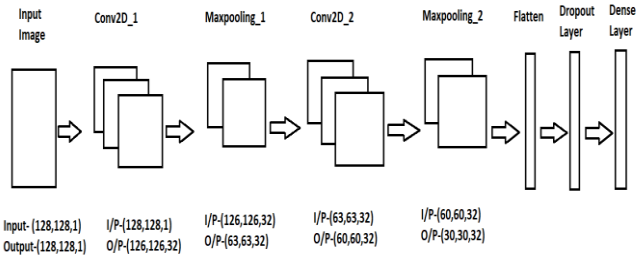

Fig. 2: CNN Architecture

3.4 Characters Recognition

The Convolutional Neural Network model has been built to separate the gesture symptoms and indicators from the historical past utilizing images. These function maps show how the CNN can recognize some of the gesture indicators in the training set that are commonly unexposed.

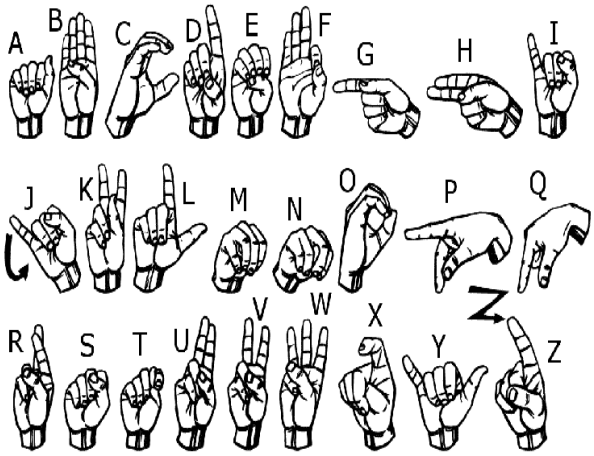

Fig. 3: The Gesture Symbols for ASL Alphabets

4. RESULTS

Using our approach, we are able to reach a model accuracy of 95.8%, which is higher than most existing research articles on American sign language. Following are some examples of our sign language recognition in user interface design of our project.

Fig. 4: GUI Screenshot 1

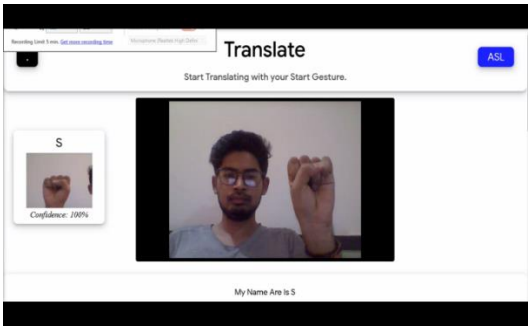

Fig. 5: GUI Screenshot 2

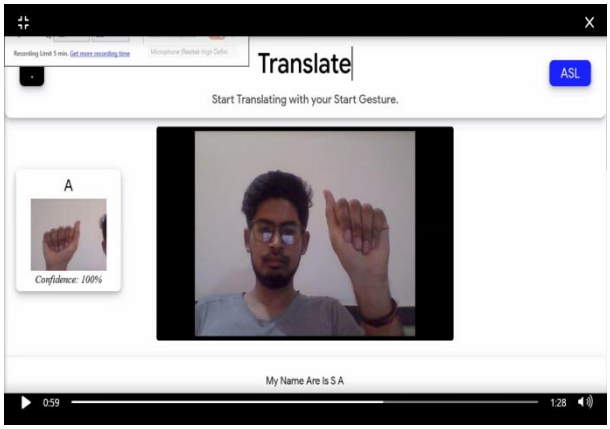

Fig. 6: GUI Screenshot 3

5. CONCLUSION

The project is a straightforward example of how CNN may be applied to computer vision difficulties. The project is able to address a portion of the Sign Language translation problem because sign languages are spoken in context rather than as finger typing languages. This project can be improved in the future in a few ways. It could be constructed as a web or mobile application enabling users to access the project more easily. By experimenting with various background removal methods, we hope to obtain improved accuracy even in the situation of complicated backgrounds. For upgrading the preprocessing to better predict gestures in low-light situations is also being considered.

6. ACKNOWLEDGEMENT

The researchers would like to thank Prof. M.P. Wankhede (Head of Department), Prof. P.M. Kamde (internal guide) and Prof. S.N. Bhosale for the endless support throughout the study.

7. REFERENCES

[1] Yang, Jie and Y. Xu. “Hidden Markov Model for Gesture Recognition.” (1994), doi: 10.21236/ada282845.

[2] Pujan Ziaie, Thomas Muller, Mary Ellen Foster , and Alois Knoll “A Naive Bayes Munich,Dept. of Informatics VI, Boltzmannstr. 3, DE-85748 Garching, Germany.

[3] Hsien-I Lin, Ming-Hsiang Hsu, Wei-Kai Chen, “Human Hand gesture recognition using a convolution neural network”, 10.1109/CoASE.2014.6899454, August 2014

[4] Junho Yim, Jeongwoo Ju, Heechul Jung, Junmo Kim, “Image Classification Using Convolutional Neural Networks With Multi-stage Feature”, Robot Intelligence Technology and Applications 3, 2015, Volume 345, ISBN : 978-3-319-16840-1

[5] S. Shahriar et al., "Real-Time American Sign Language Recognition Using Skin Segmentation and Image Category Classification with Convolutional Neural Network and Deep Learning," TENCON 2018 - 2018 IEEE Region 10 Conference, 2018, pp. 1168-1171, doi: 10.1109/TENCON.2018.8650524.

[6] D. Kelly, J. Mc Donald and C. Markham, "Weakly Supervised Training of a Sign Language Recognition System Using Multiple Instance Learning Density Matrices," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, no. 2, pp. 526-541, April 2011, doi: 10.1109/TSMCB.2010.2065802.

[7] López-Ludeña, Verónica & San-Segundo, Rubén & Martin, Raquel & Sanchez, David & Garcia, Adolfo. (2011). Evaluating a Speech Communication System for Deaf People. IEEE Latin America Transactions - IEEE LAT AM TRANS. 9. 565-570. 10.1109/TLA.2011.5993744.

[8] E. Abraham, A. Nayak and A. Iqbal, "Real-Time Translation of Indian Sign Language using LSTM," 2019 Global Conference for Advancement in Technology (GCAT), BENGALURU, India, 2019, pp. 1-5, doi: 10.1109/GCAT47503.2019.8978343.

[9] Arsan, Taner & Ulgen, Oguz. (2015). Sign Language Converter. International Journal of Computer Science & Engineering Survey. 6. 39-51. 10.5121/ijcses.2015.6403.

[10] V. N. T. Truong, C. Yang and Q. Tran, "A translator for American sign language to text and speech," 2016 IEEE 5th Global Conference on Consumer Electronics, 2016, pp. 1-2, doi: 10.1109/GCCE.2016.7800427.

