2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) | 978-1-6654-2829-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAECA52838.2021.9675751

Signtalk: Sign Language to Text and Speech Conversion

C.Uma Bharathi
Department of ECE
Kumaraguru College of Techonology
Coimbtore,India
uma.17ec@kct.ac.in

G.Ragavi
Department of ECE
Kumaraguru College of Techonology
Coimbtore,India
ragavi.17ec@kct.ac.in

K.Karthika
Department of ECE
Kumaraguru College of Techonology
Coimbtore,India
karthika.k.ece@kct.ac.in

***Abstract*— People with hearing and speech impairments are facing lots of difficulties while communicating with the public. The proposed work provides a helping hand for hearing/speech-impaired and even blind people to communicate with others. For a sign input, the proposed model provides a speech/text output, thus providing a user-friendly platform for users. Being a minority, the sign language used by them is not known to most people. So, the idea proposed is a system which converts American Sign Language (ASL) to text and speech output. This work uses convolutional neural networks (CNN) to extract efficient hand features to identify the hand gestures according to ASL. The proposed model offers an accuracy of 88%. Hence, this system helps in recognizing the hand gestures of special people and converting them to text and speech to communicate more effectively with normal people.**

Key words: CNN, ASL, sign language, text, speech, hand gesture, GTTS

I. INTRODUCTION

Humans interact among each other and express their thoughts using vocal sounds in the form of spoken languages, facial expressions, hand movement and body language. The spoken language remains the most effective tool for communication. But unfortunately, for people with disabilities such as hearing impairment or speech impairment, spoken language does not serve this purpose. According to the census taken in India in the year 2011, 2.1% of the total population suffer from some form of disability. Among those disabled persons, 7.5 % have speech disability and 18.9 % have hearing impairment. So, a significant percentage of the Indian population needs an effective tool to bridge the communication deficit. The practical solution to this problem is to learn sign language. Though the use of sign language has been around all over the world for many years, most people still don't understand sign language. This makes it difficult for people with hearing impairments and speech disorders to effectively communicate with others without a translator for sign language [1]. Many researchers have contributed to overcoming this difficulty by developing systems that can convert sign language to text or sign language to speech.

In [2], the authors have built a system called HandTalk that comprises both software and hardware components. The hardware components include a Bluetooth data acquisition module, gloves fixed with a flex sensor, and a smart phone. The associated software performs the function of converting sensor data to recognized signs. MIDlet is written in Java 2 Micro Edition to do this. It runs on a Personal Digital Assistant (PDA). The user's movements are captured by the flex sensor in the gloves, and the Bluetooth data gathering module quantizes the data and wirelessly feeds it to a mobile phone or PDA through Bluetooth technology. The Nokia N95 is used in the prototype, although the system may be operated on a variety of phones and PDAs. The Hand Talk MIDlet uses a cached database that links bend measures to ASL signals to transform the quantized data into text. The text is then converted to speech using Nuance Communications' Real-Speak program.

The authors of [6] suggested a system that is intended for the differently abled section of society to assist in the conversion of sign language into a more human comprehensible form, such as written communications. Their main objective is to create a low-cost wired interactive glove that can connect to a computer running MATLAB or Octave software. Hall Effect sensors and accelerometers are used to map the orientation of the hand and fingers. The data is then sent to a computer through UART to MATLAB script utilizing ARQ. In [3], authors build a skin model to extract an image's handout and then apply a binary threshold to the entire image.After obtaining the threshold image, they calibrate it about the principal axis to center the image around it. They input this image to a convolutional neural network model to train and predict the outputs. They have trained their model over seven hand gestures, and using their model, they produce an accuracy of around 95% for those 7 gestures. In [7], Indian sign language is converted to text. Their model got an accuracy of 87.69. In [8], a recognition model is built using a hidden Markov model classifier and a vocabulary of 30 words, and they achieve an error rate of 10.90%. In [9], authors have achieved an average accuracy of 86% for 41 static gestures in Japanese sign language. A depth sensor map is used to achieve an accuracy of 85.49% for new signers [10]. CNN is used for feature recognition [3], [5], [10], [12].

ASL is the most-used sign language among the disabled. ASL is generally favoured as the communication tool for deaf and dumb people [2]. For a hand gesture recognition system, the number of datasets of an image is widely large, hence it is difficult to extract the required features of the image [11]. This work intends to help healthy individuals understand people with speech impairment without the need for a translator. The significant contribution of this work is to develop a vision-based system to convert ASL gestures into text and speech. The proposed work utilizes CNN algorithm for feature recognition and Google text to speech engine for text to speech conversion. This paper is sectioned as follows. Section II discusses the process involved in developing the proposed model. Section III discusses the results of the model and Section IV concludes the paper.

II. METHODS AND MATERIALS

The primary objective of this work is to create a vision-based application that can capture sign language and translate it into text and eventually, into speech. This enables signers to easily communicate with non-signers without any hinderance. The proposed model captures images from video sequences and those will be used as an

978-1-6654-2829-3/21/$31.00 ©2021 IEEE

input. The Python software will recognize the image and identify the text output. To implement a good hand gesture recognition system, a large training database is usually required. Hence, the CNN algorithm is utilized for recognizing features. Further, for the conversion of text into speech, the Google text into speech engine is used. The process involved in the proposed system is depicted in Fig.1.

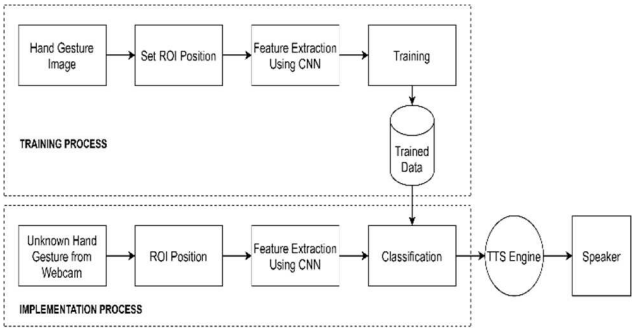
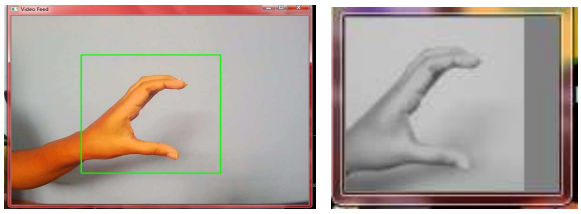


Fig.1. Proposed Model - Sign to Text and Speech Conversion

A. Dataset

Datasets in the form of raw images are not found. Available datasets are in the form of RGB values. Hence, it is decided to create their own data set. The Open computer vision (OpenCV) library is used to produce its own dataset. Firstly, around 1000 images of each of the symbols in ASL for training purposes are captured and around 300 images per symbol (i.e., 70:30 ratio) are used for testing purposes. First, each frame shown by the webcam of the machine is captured. A total of 12 labels were assigned. Each label has 1000 images, therefore a total of 12000 images are used for the process. In each frame, a region of interest (ROI) is defined and it is denoted by a green boundary as depicted in Fig.2 (a).



(a) (b)

Fig.2. (a) ROI (b) Gray scale image

ROI where RGB is extracted from this whole image and converted into a grayscale image as shown in Fig.2 (b). A Gaussian blur filter is used to extract various features of the image. It is highly important to mitigate the computational effort needed to process images taken from the camera. In the proposed system, the person with a disability should provide a gesture or sign image to the system. The proposed work is a vision-based system. Since all the signs are made by bare hands, there is no need to use artificial devices for interaction. Various symbols along with their meanings are shown in Fig.3.

Sign	Meaning
	Hello
	How are you
	I am hungry
	I am thirsty
	Please help
	Alphabet C
	Alphabet A
	Alphabet B
	Alphabet I
	Alphabet L
	Alphabet Y
	Alphabet W

Fig.3. Various Signs and their Meanings

B. Training

In the training process, a webcam is employed to capture a video frame. A particular region of the captured video frame is identified and designated as a Region of Interest (ROI). This process eliminates the unwanted space from the video frame. Noises occur either due to light effects or due to the image background. This can be eliminated using a median filter (Gaussian Blur). This process is followed by the gray scale image conversion of all the images. Then, the hand gesture pictures are taken for ASL. The hand gesture features are extracted using CNN and it undergoes training.

C. Implementation

In the implementation process, the hand gestures of the user are captured by a webcam. Among the multiple ways to perform feature extraction, CNN is the simplest in the deep learning domain. Further, CNN is also competent in extracting potential features from a large scale of diverse images. Hence, this model utilizes a Convolution Neural Network (CNN) to withdraw the feature vector from a particular video frame. Feature values that are extracted from the images are kept in a data file. These extracted features are then classified. The data from this process is compared with trained data. Based on the accuracy of both the data, the message is displayed. Among the numerous networks of CNN, "Convnet" is by far the most well-established because of its effectiveness.

D. Google Text to Speech Engine (GTTS)

GTTS is a Python library that converts text into audio. The play sound module is used to play audio files. With this module, it is possible to play a sound file with a single code line. This software accepts text as input from the user. Then, utilizing natural language processors, it understands the language that has been used and converts it into speech.

E. Play Sound

The play sound module is a cross platform module that can play audio files. The Play Sound function plays a sound mentioned by a resource, file name or system event. This function will create an audio file (.mp3) that has the Google voice saying whatever text is being passed in. Further, the file will be saved to the same directory.

F. Gesture Classification

Gaussian blur filter to the frame taken with OpenCV to get the processed image after feature extraction is applied. The processed image is passed to the CNN model for prediction and based on the probability value, the letter is printed and taken into consideration. The proposed approach uses five layers to predict the final symbol of the user.

G. Training and Testing

All the input images (RGB) are converted into grayscale and applied with Gaussian blur to remove unnecessary noise. The images are resized to 96 x 96. The input images are feed to the proposed model for training and testing after applying all the operations mentioned above. The prediction layer estimates how likely the image will fall under one of the classes. So, the output is normalized between 0 and 1 and such that the sum of each value in each class sums to 1. This was achieved using the SoftMax function. At first, the output of the prediction layer will be somewhat far from the actual value. To make it better, the networks were trained using labeled data. TensorFlow has an inbuilt function to calculate the cross entropy. The cross-entropy function was found and optimized using a gradient descent optimizer called the Adam Optimizer.

III. Results And Discussion

This methodology discusses the challenges involved in the practical applications of hand gesture recognition. It analyzes its performance in terms of accuracy and impact. This paper discusses the recognition of ASL from real time hand gestures. Hand signs for each gesture or alphabet are collected with the help of the webcam. Firstly, the still image of hand sign gesture is obtained from a video frame. Furthermore, a CNN operation has been carried out to get more information on the image's features.

The training utilizes the Adam optimizer with a categorical cross-entropy loss function. The proposed model is provided with 12 gestures, each gesture is having 1000 images. So, in the training process 12000 images are used. 300 images per gesture i.e., 3600 images are used for testing process. As shown in Fig.4, the computer's web cam is used as an input device which will capture the real time image and display the output text and the system's speaker provides us with speech output with respect to the text output.

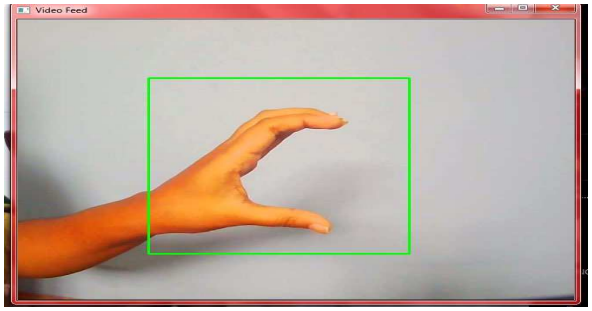


Fig.4. Real time Image capturing.

The model has 15 epochs and a learning rate of 0.001. Accuracy achieved by the proposed model is 88%. The proposed model exhibits a low loss of 0.2097 with limited overfitting of 0.8 after 15 epochs. The accuracy achieved by this model is better than most of the current research work on American sign language. The Fig.5 depicts the sample output of the proposed system.

TABLE I Comparison of Proposed Work with Existing Works

Ref. No.	Significance	Accuracy
[7]	Indian sign language converter using CNN	87.69%
[8]	Sign language recognition implemented with novel vision-based features	87.56%
[9]	Recognition of Japanese fingerspelling using a classification tree and machine learning	86%
[10]	Real time sign language fingerspelling recognition using CNN	85.49%
Proposed work	Real time sign language to text and speech conversion	88%

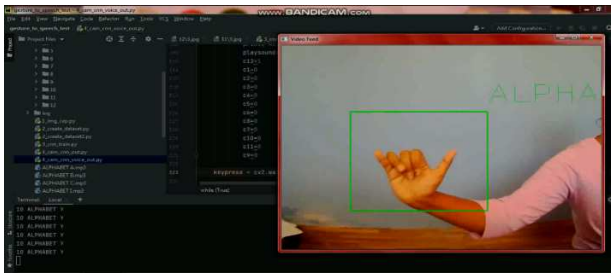


Fig.5. Sample Result-Sign Language to Text and Speech

The findings of the experiments suggest that this technique is more effective at anticipating real-time static gestures, and it may also be used as a first step toward bridging the communication gap between people with hearing/speech impairments and the public. This will also be able to support advanced sign language that includes hand motions. The comparison of the proposed work with previous works is shown in Table I. This will also accommodate more advanced sign language involving gestures. Most of the previous work uses a background subtraction algorithm which segments moving objects from the main frame by detecting the background and input images.

The proposed model did not use any background subtraction algorithm as the dataset was created with stable background. While some models that existed before did use it. Most of the research papers focus on using Kinect devices, but the main aim of the proposed work was to create a project which could be used with readily available resources. A sensor like Kinect devices is not only readily available but also expensive for most of the audience to buy and the model uses a normal webcam and speaker of the laptop, hence it is a major advantage.

IV. CONCLUSION

For people with hearing impairments and speech difficulties, the proposed system offers a ray of hope. It acts as a bridge between the general society, which is ignorant of sign language, and the people with disabilities, who solely depend on sign language for their communication. The proposed work uses no additional hardware, thus making it much simpler to use than the alternative existing work. Many applications utilize computer vision systems to gather information. In this report, a functional real-time vision-based American sign language recognition system for people with hearing impairments and speech difficulties. The model achieved final accuracy of 88% with the provided dataset. The predictions were improved after implementing five hidden layers in which to verify and predict symbols that are more like each other. This way, the proposed model can detect almost all the symbols if they are shown properly, there is no noise in the background, and lighting is adequate.

REFERENCES

[1] M. C. Su, C. Y. Chen, S. Y. Su, C. H. Chou, H. F. Hsiu, and Y. C. Wang, “Portable communication aid for deaf-blind people,” Comput. Control Eng. J., vol. 12, no. 1, pp. 37–43, 2001, doi: 10.1049/cce:20010106.

[2] D. K. Sarji, "HandTalk: Assistive Technology for the Deaf," in Computer, vol. 41, no. 7, pp. 84-86, July 2008, doi: 10.1109/MC.2008.226.

[3] H. Lin, M. Hsu and W. Chen, "Human hand gesture recognition using a convolution neural network," 2014 IEEE International Conference on Automation Science and Engineering (CASE), 2014, pp. 1038-1043, doi: 10.1109/CoASE.2014.6899454.

[4] S. Konadath, C. Suma, G. Jayaram, M. Sandeep, G. Mahima, and P. S. Shreyank, “PREVALENCE OF COMMUNICATION DISORDERS IN A RURAL POPULATION OF INDIA,” J. Hear. Sci., vol. 3, no. 2, pp. 41–49, 2013.

[5] L. Y. Bin, G. Y. Huann and L. K. Yun, "Study of Convolutional Neural Network in Recognizing Static American Sign Language," 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2019, pp. 41-45, doi: 10.1109/ICSIPA45851.2019.8977767.

[6] T. Chouhan, A. Panse, A. K. Voona and S. M. Sameer, "Smart glove with gesture recognition ability for the hearing and speech impaired," 2014 IEEE Global Humanitarian Technology Conference - South Asia Satellite (GHTC-SAS), 2014, pp. 105-110, doi: 10.1109/GHTC-SAS.2014.6967567.

[7] N. Intwala, A. Banerjee, Meenakshi and N. Gala, "Indian Sign Language converter using Convolutional Neural Networks," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033667.

[8] M. M. Zaki and S. I. Shaheen, “Sign language recognition using a combination of new vision based features,” Pattern Recognit. Lett., vol. 32, no. 4, pp. 572–577, 2011, doi: https://doi.org/10.1016/j.patrec.2010.11.013.

[9] N. Mukai, N. Harada and Y. Chang, "Japanese Fingerspelling Recognition Based on Classification Tree and Machine Learning," 2017 Nicograph International (NicoInt), 2017, pp. 19-24, doi: 10.1109/NICOInt.2017.9.

[10] B. Kang, S. Tripathi and T. Q. Nguyen, "Real-time sign language fingerspelling recognition using convolutional neural networks from depth map," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015, pp. 136-140, doi: 10.1109/ACPR.2015.7486481.

[11] H. V. Guda, S. Guntur, G. P. M, K. Gupta, P. Volam, and S. P V, “Hardware Implementation of Sign Language to Text Converter Using Deep Neural Networks,” SSRN Electron. J., 2020, doi: 10.2139/ssrn.3576354.

[12] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, “Sign Language Recognition Using Convolutional Neural Networks,” in Computer Vision - ECCV 2014 Workshops, 2015, pp. 572–578.

[13] S. Rajaganapathy, B. Aravind, B. Keerthana, and M. Sivagami, “Conversation of Sign Language to Speech with Human Gestures,” Procedia Comput. Sci., vol. 50, pp. 10–15, 2015, doi: https://doi.org/10.1016/j.procs.2015.04.004.

[14] P. Vijayalakshmi and M. Aarthi, "Sign language to speech conversion," 2016 International Conference on Recent Trends in Information Technology (ICRTIT), 2016, pp. 1-6, doi: 10.1109/ICRTIT.2016.7569545.