HandSpeak: Gesture Recognition System using Machine Learning

Rishav Anand
Department of Computer Science and Engineering
Galgotias University, Greater Noida, U.P., India
19rishav02@gmail.com

Tushar Gupta
Department of Computer Science and Engineering
Galgotias University, Greater Noida, U.P., India
tg885171@gmail.com

Arpan Kumari
Department of Computer Science and Engineering
Galgotias University, Greater Noida, U.P., India
arpan.kumari@galgotiasuniversity.edu.in

Hand Speak, is named as about the people who are challenged by the vision, a speech limitation. Some of the previous studies have recognized sign language. Still, it takes high-level, costly equipment and sensors but in this generation, this new era of AI-based (artificial intelligence) techniques has easily overcome these situations. In this environment where, cameras can take videos, and images easily so now this paper will provide overview how the trained model presents a cost-effective technique to detect American sign language using an image dataset. During this work, users use not so sophisticated device just only the webcam to take images of hand movements and the computer and after the system predicts and shows the name of the gesture. The predicted image taken goes through multiple processing operations such as dilatation, mask operation, and conversion to grayscale, among other computer-imaginative and predictive techniques. The results confirmed that they were beneficial for the community and were of good use in different sectors. Furthermore, we have supported four supporting applications for our system to show how it will be applied in the real-life world.

Keywords: ASL (American Sign Language), Deaf - Non vocal persons, hand gestures, CNNs.

SECTION I.

I. INTRODUCTION

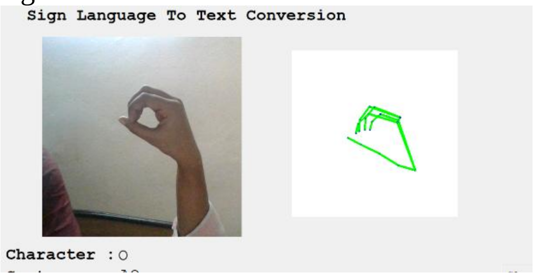
This paper is how the system can link between individuals with auditory impairments and the computer. However, it remains unfamiliar and not understood by broader public, those who are speech and hearing-impaired often find themselves relying on human translators. Unfortunately, access to human interpreters can be inconsistent and costly. The best of action is to create a communication system that can reliably read and translate sign language convert into a more accessible format. These society’s communication gap will be reduced by this cutting- edge technology. Image Recognition is a dynamically developing area where it uses artificial intelligence nowadays. In the real-time world, human work is successfully done by supporting ANN including machine image recognition systems, including healthcare [1,2], monitoring [3], and various others. Convolutional neural networks are the one of the tools that has been widely used for these works [4]. CNN takes input as an image the network processes it in multiple steps that summarize the input and retain most critical features for later stages in the network. Apply filters called as kernel that slide over the image to detect patterns like edges, and curves now reduce the computation and prevent overlifting on the extracted (output) value from the set by using pooling layers or max-pooling [5] by filtering the informational features. Then the operation processes extracted features to predict output as gesture (‘A, B……., Z’ in American Sign Language that often uses SoftMax for multi-classification [6].

II. AUTOMATED COMMUNICATION SYSTEM

This is a model that can recognize sign gestures. The gesture usually given with the hand and fingers movement. Once the sign (input) captured by the camera and feeds in the pre-processing interpreter model.

The CNN model is trained with the labeled of hand gestures dataset of more than 4500 input of different letters (A, B…Z) that increases the prediction accuracy.

CNN model extracts hand shape, orientation, and finger positions and it predicts the most likely sign and result will be on the screen.


Sign Language To Text Conversion

Character : O

Fig.1 (Love's one-hand gesture has shown that is detected and pre-processing actions on and predicted output shown as character as 'O')

SECTION II.

RELATED WORK

With the advancement of technology, Materials and Methods

Python version 3.12.3, global programming environment, and were utilizing "the Keras along with TensorFlow libraries" for developing both its design and training, and here, CNN convolutional neural network architecture implemented to construct the model—OpenCV - Python for capturing webcam, detecting hands where mediapipe activate hand landmark for recognizing signs. The dataset used for training included upwards of 4500 images each measuring 200 *200 pixels, were collected to represent the 26 characters of American Sign Language alphabets that have been driven as downloaded from the online platform Kaggle.

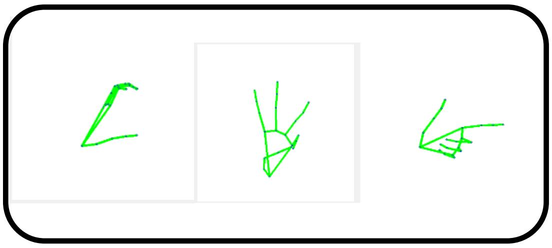


Fig2. Sign language characters dataset from the platform Kaggle.

Operation of the data will be processed set on a data module when the test data set is analyzed in the process of training, The data initial dataset was systematized and reduced or resized the image that goes through in the CNN model. As a result showing the prediction ('A', 'B', … 'Z') and the text displayed via CVZone.

SECTION III.

DESIGN

System Architecture

The structure of the device where include input acquisition as image through a web cam, after it passed through the pre processed module where the image is resized and cropped focused on the hand landmarks extracted using frameworks Mediapipe that gives a better segmentaion .

The pre-processed images are fed into the CNN which consists of layers:

- To extract spatial features, use convolutional layers.

- Reducing dimension by pooling layers

- fully connected layers are utilized to decode the features representation from the processed image

- A gesture is classified into one of the predefined categories (such as A-Z) using the Softmax output layer [7] Fig 3.

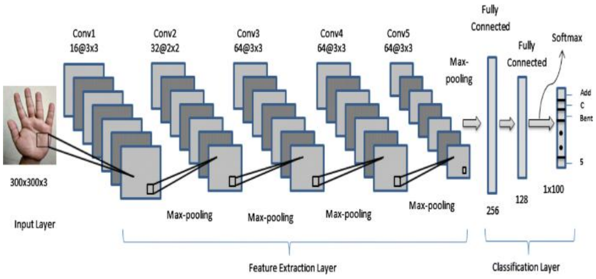


Fig3. Image processed through the CNN layers

The numerical output is interpreted into a corresponding character that predicts in single letter and allows user to see the webcam feed and the predicted gesture in real-time.

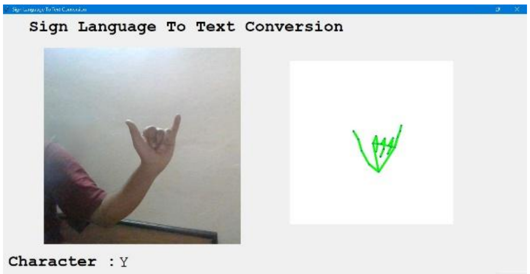
SECTION IV.

DEVELOPING THE SYSTEM

A. PROCEDURE

The proposed methodology for the SLR is designed so, system can observe the static images of hand gestures corresponding to the alphabets (A-Z) using a CNN model (Convolutional Neural Network).

- Data Collection: A dataset consists of hand gesture that represents the letters that is collected by the online platform Kaggle and labelled according to the corresponding alphabets.
- Image Preprocessing: Captured images are resized and normalized the pixel values for improved the model performance. Filtering and segmentation are applied using the mediapipe to capture and extract the hand region.
- Development of Convolutional Neural Networks: Here it creates the two convolutional layers for responsible of pooling that extract features while at this time ReLU activates for efficiency in training the large neural networks that work as extraction of complex patterns from gesture images. Fully connected layers classify all the features to decide which sign (A–Z) the image represents.
- Training and Tests of the Model: The CNN is trained using the preprocessed dataset. Once the model has been trained now it should detect the gesture from the image.
- Demonstration Module for Real-Time Gesture Recognition: A webcam of the computer as in online real-time captures the hand gesture, then each frame is pre-processed and passed through the trained CNN and the processing result is presented in the webcam feed Fig 4.



Sign Language To Text Conversion

Character : Y

Fig 4. Visualization of the interface showing the predictable output as ‘Y’

B. METHODOLOGY

Convolutional Neural Networks (CNNs) have efficiently worked in sign language recognition by capitalizing on their ability to autonomously learn and extract key features from visual data enabling accurate gesture and sign classification. A CNN as a machine learning system have set of rules that can take receives an input image, prioritizes to an object, and then try to differentiate between one item and others by extracting capabilities from the pixels.

CNN consists of three components, which could be: an input layer that is a grayscale photograph, an output layer that is the binary o multiclass labels, and third hidden layers that include a convolution layer, RELU, pooling layers, and finally an artificial neural community to perform the type. Let's look at CNN’s structure.[8]

In CNN structure, we start with an input as image or photo, which we then want to convert it into pixels. Subsequently the dense layers categorize the image and Output layer provides the ultimate prediction that is gesture or sign.

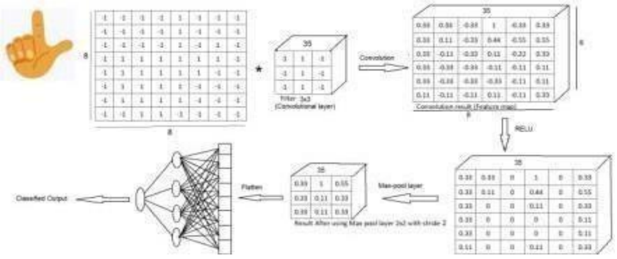


Fig 5. CNN Architecture

C. RESULT

Visualization and Identifying Unknown Images

For the purpose of this activity, we must check the effectiveness of the model for accurately identify the characters. So, we have given the input 10 real time hand gesture to the webcam. **Fig 6.**

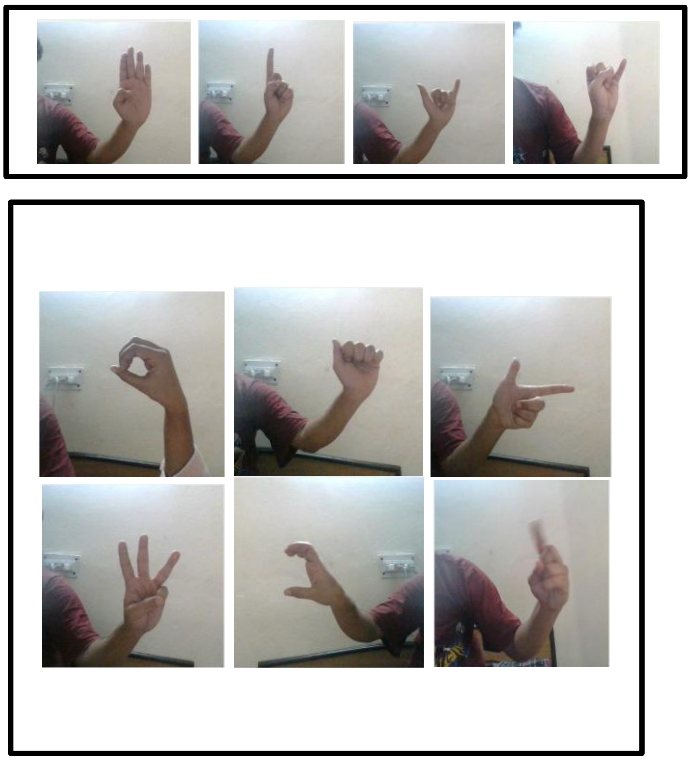
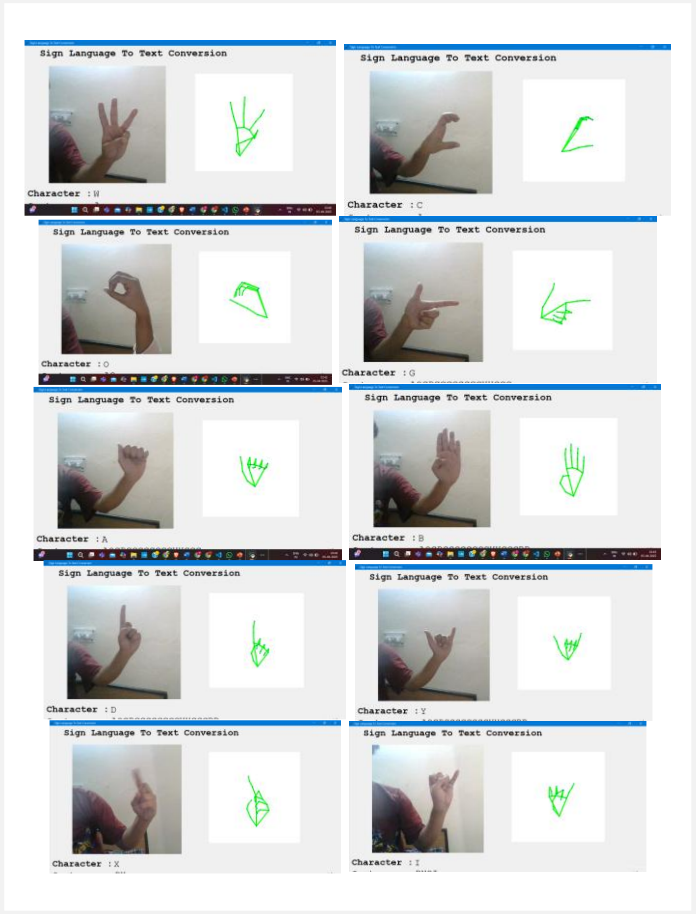


Fig 6. Set of real-time images for testing

The Sign recognition System outcomes displayed in **Fig 7** were acquired using the chosen photos.



Sign Language To Text Conversion
Character : W

Sign Language To Text Conversion
Character : C

Sign Language To Text Conversion
Character : O

Sign Language To Text Conversion
Character : G

Sign Language To Text Conversion
Character : A

Sign Language To Text Conversion
Character : B

Sign Language To Text Conversion
Character : D

Sign Language To Text Conversion
Character : Y

Sign Language To Text Conversion
Character : X

Sign Language To Text Conversion
Character : I

Fig 7. Sample of recognition results

Here The Presented photo shows that 8 out of 10 analyzed signs were correctly recognized in the **Fig 7** and 2 output is incorrect for the letter 'J' and 'Z' because as web cam accept static photo but for both these letters gestures are in dynamic sign so it provides inaccurate prediction for them.

D. DISCUSSION

Fig 7. Shows that 8 out of 10 is accurately predicted by this model. The drawback is that in this web - based application does not captured or detect dynamic hand gestures. Kamil Kozyra [8] has the accuracy of 99% where in his study set of 50,000 photos was used, where 40,000 of these were taken as training material and 10,000 were used for checking the learning results. The dataset proved to be a notable level of effectiveness that reached 99% and 24 out of 40 images were identified.[8]

In future we think to make more efficiently and accurate recognition so that it gets better result. A solution came up that we can use hybrid approach of CNN With RNN (Recurrent Neural Networks), particularly LSTMs enabling them to capture the movements in sign language gestures. [9] shows

CNN Feature Extraction:

The CNN processes the input images (or frames) to extract spatial features, such as handshapes and other visual cues.

RNN Processing:

The extracted features are then fed into an RNN (e.g., LSTM), which handles the sequential feature data over time, capturing the time-dependent variations sign language gestures.

Output:

The RNN generates result, classifying the sign language gesture or word.

SECTION V.

CONCLUSION AND FUTURE SCOPES

In this study, we have demonstrated for sign language recognition. We have dedicated on the development of CNN-based model capable of classifies the gestures from a static image with real time efficiency. For training purpose, a set of 4500+ photos taken as training model and for testing purpose approximately 100+ photos to validate the results. The analysis of the training dataset was observed to achieved for 87%, where the letter 'J' and 'Z' not properly working in this model because of its dynamic movement on the gesture over time. In Python environment with the help of the Keras library that were utilized to construct and trained the model. Based on the web-cam was accessed by opencv-python to test the images. It supports detection of sign alphabet gestures from an input image taken by an individual. The system has been tested on the photos that does not belongs to the training set. This led to, 28 out of 38 were accurately predicted because the quality of the photos and particularly in clarity and lightning plays a vital role that responsible for negative impact on character detection. The dataset applied in this study has been one-handed gestures so it varies when the greater variation will be tested.

This system is expected to be explored in the field of human computer interaction. Future improvements to this method should recognize a wider range of signal languages, promoting inclusion and this facilitates access for hearing-impaired individuals. Furthermore, continuous advancements may improve the effectiveness and precision of hand sign recognition that helps ongoing efforts to develop accurate and accessible sign language and be helpful in the real time translation, communication and accessibility in public places like airport, hospitals and many more different sectors.

SECTION VI.

REFERENCES

[1] Mendels, D., Dortet, L., Emeraud, C., Oueslati, S., Girlich, D., Ronat, J., Bernabeu, S., Bahi, S., Atkinson, G. J., & Naas, T. (2021). Using artificial intelligence to improve COVID-19 rapid diagnostic test result interpretation. *Proceedings of the National Academy of Sciences*, *118*(12), e2019893118. https://doi.org/10.1073/pnas.2019893118

[2] Thurzo, A., Kosnáčová, H. S., Kurilová, V., Kosmeľ, S., Beňuš, R., Moravanský, N., Kováč, P., Kuracinová, K. M., Palkovič, M., & Varga, I. (2021). Use of Advanced Artificial Intelligence in Forensic Medicine, Forensic Anthropology and Clinical Anatomy. *Healthcare*, *9*(11), 1545. https://doi.org/10.3390/healthcare9111545

[3] Nguyen, M. T., Truong, L. H., Tran, T. T., & Chien, C. (2020). Artificial intelligence based data processing algorithm for video surveillance to empower industry 3.5. *Computers & Industrial Engineering*, *148*, 106671. https://doi.org/10.1016/j.cie.2020.106671

[4] Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 6999-7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[5] Bieder, F., Sandkühler, R., & Cattin, P. C. (2021). Comparison of Methods Generalizing Max- and Average-Pooling. *ArXiv*. https://arxiv.org/abs/2103.01746

[6] M. A. Hussain and T. -H. Tsai, "An Efficient and Fast Softmax Hardware Architecture (EFSHA) for Deep Neural Networks," 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS), Washington DC, DC, USA, 2021, pp. 1-4, doi: 10.1109/AICAS51828.2021.9458541.

[7] Wadhawan, A., Kumar, P. Deep learning-based sign language recognition system for static signs. *Neural Comput & Applic* **32**, 7957–7968 (2020). https://doi.org/10.1007/s00521-019-04691-y

[8] Sanket Bankar1, Tushar Kadam2, Vedant Korhale3, Mrs.A.A.Kulkarni

[9] Kozyra, K., Trzyniec, K., Popardowski, E., & Stachurska, M. (2021). Application for Recognizing Sign Language Gestures Based on an Artificial Neural Network. *Sensors*, *22*(24), 9864. https://doi.org/10.3390/s22249864

[10] Ko, Sang-Ki & Son, Jae & Jung, Hyedong. (2018). Sign language recognition with recurrent neural network using human keypoint detection. RACS '18: Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems. 326-328. 10.1145/3264746.3264805.

[11] A. Gupta, A. Sawan, S. Singh and S. Kumari, "Dynamic Sign Language Recognition with Hybrid CNN-LSTM and 1D Convolutional Layers," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024, pp. 1-6, doi:10.1109/ICRITO61523.2024.1052233