International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT23902103

American Sign Language Recognition and Generation : A CNN-based Approach

Pusti Sheth[1], Ronik Dedhia[1], Akshit Chheda[1], Dr. Vinaya Sawant[2]

[1]Students, Department of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India

[2]Head of Department, Department of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 10 July 2023

Published: 28 July 2023

Publication Issue

Volume 9, Issue 4

July-August-2023

Page Number

241-250

ABSTRACT

Although not a global language, sign language is an essential tool for the deaf community. Communication between these communities and hearing population is severely hampered by this, as human-based interpretation can be both costly and time-consuming. In this paper, we present a real-time American Sign Language (ASL) generation and recognition system that makes use of Convolutional Neural Networks and deep learning (CNNs). Despite differences in lighting, skin tones, and backdrops, our technology is capable of correctly identifying and generating ASL signs. We trained our model on a large dataset of ASL signs in order to obtain a high level of accuracy. Our findings show that, with accuracy rates of 98.53% and 98.84%, respectively, our system achieves high accuracy rates in both training and validation. Our approach uses the advantages of CNNs to accomplish quick and precise recognition of individual letters and words, making it particularly effective for sign fingerspelling recognition. We believe that our technology has the ability to transform communication between the hearing community and the deaf and hard-of-hearing communities by providing a dependable and cost-effective way of sign language interpretation. Our method could help people who use sign language communicate more easily and live better in a range of environments, including schools, hospitals, and public places.

Keywords : American Sign Language (ASL), Convolutional Neural Networks (CNNs), Deep learning, Accuracy rates

Copyright © 2023 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution **4.0 International License (CC BY-NC 4.0)** which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

I. INTRODUCTION

Effective communication is essential for survival in society, but it can be difficult for individuals who are deaf-mute to communicate with those who can hear and speak. This is largely due to a lack of understanding of Sign Language, which is the primary means of communication for deaf-mute individuals. As a result, communication between deaf-mute individuals and others can be challenging. Most existing messaging systems do not have features for detecting or generating Sign Language, which creates significant barriers for individuals with disabilities who rely on this mode of communication. This lack of accessibility limits the online presence of individuals with disabilities, preventing them from fully participating in the digital world that many of us take for granted. Sign Language Recognition technology is a significant advancement that can greatly benefit individuals who are deaf-mute. However, the cost of acquiring the necessary devices can be a barrier to widespread adoption. It is crucial to develop affordable solutions to enable the commercialization of this technology. The primary goal of this technology is to promote social inclusion and integrate people with disabilities into society. Computer vision can also be utilized to further support this cause.

The world has a huge variety of languages. With the aid of software like Google Translate, nearly all of the most common languages can be translated in real-time. The software is primarily based on an NLP algorithm that receives input from the text in one language and generates its translation in text. Some software uses a microphone and speakers to translate a message in real-time rather than requiring the user to type it in. This is quite effective with spoken languages. What about the 1 billion people who use sign language to communicate because they are unable to hear or speak? The usage of "Sign Language" is the sole viable communication method for those with disabilities. They can only communicate in their small environment through sign language. Because of this restriction, they are unable to interact with others and express their emotions, thoughts, and potential. According to research, 90% of people who are not deaf or mute have trouble comprehending or communicating with sign language, which further isolates disabled people from the general population. With 2.4 billion active mobile phone subscribers, SMS text messaging is a very common means of communication. Chat rooms can still be challenging for persons with disabilities to utilize, despite technological developments. It's critical to foster an inclusive culture and build technological solutions that can help people with impairments.

For the deaf community, American Sign Language (ASL) is a vital form of communication. However, the limited number of ASL speakers, estimated to be between 250,000 to 500,000, presents significant challenges for deaf individuals seeking to communicate with the wider population. While written communication is an option, it can be time-consuming, impersonal, and unworkable in urgent or emergency circumstances.

In general, ASL recognition and Sign Language Generation technologies are essential for fostering social inclusion and communication for those with hearing loss. We can ensure that everyone may fully engage in the digital world and beyond by making it possible for successful communication between the deaf and hearing communities.

II. RELATED WORK

A review of the relevant literature revealed that many techniques and algorithms have been researched to deal with the issue of sign detection in videos and photos.

Computer vision issues with ASL recognition are nothing new. Researchers have utilized a number of classifiers over the past 20 years, which we may loosely

divide into three categories: linear classifiers, neural networks, and Bayesian networks [1–7].

A system using a collection of 10,000 images of sign language and 40 frequently used terms were suggested in a research report [8]. The method combines Faster R- CNN with an incorporated RPN module to locate the hand regions in video frames, which enhances accuracy in comparison to single-stage target identification techniques like YOLO. When compared to Fast R-CNN, the detection accuracy of the Faster R-CNN in the paper increased from 89.0% to 91.7%. The framework for detecting sign languages uses long and short-term memory (LSTM) coding and decoding networks for language picture sequences, as well as a 3D CNN for feature extraction. In this paper, an extraction method for the RGB sign language image or video challenge is developed by combining the hand-locating network, 3D CNN feature extraction network, and LSTM encoding and decoding. A 99% identification rate for the common vocabulary dataset was attained by the article.

Using a dataset of 50 samples that contained all alphabets and numbers, M. Geetha and U.C. Manjusha developed a vision-based recognition system for Indian Sign Language characters and numerals in their work [9]. The system required assessing the sign gesture's region of interest and eradicating the boundary. Using Maximum Curvature Points (MCPs) as control points, the boundary was then converted into a B-spline curve. To extract features, the B-spline curve underwent a number of smoothing procedures. The images were categorized using a support vector machine, and its accuracy was 90.00%.

In a study detailed in [10], pictures were collected with a green background and processed using a low-cost approach. This made it simple to subtract the colour green from the RGB colour space, producing a black-and-white image. The sign language used was in the Sinhala alphabet. With the centroid method, which can perfectly match input motions with a database independent of the size and placement of the hand, the suggested method maps the indications. The prototype created using this methodology had a 92% success rate in recognizing sign motions.

The fully connected layer, also known as the ANN, is utilized for classification whereas the CNN is in charge of feature extraction. In a research [11], the suggested system had an error rate of 8.30% and a 91.70% accuracy rate. In a different study, J. Huang used Kinect to build a custom dataset with 25 common sign language vocabularies. In this work, a 3D CNN with 3D kernels was employed. Important channels including colour-b, colour-g, colour-r, body skeleton and depth were included in the input. The system's accuracy was on average 94.2%.

High accuracy has also been attained by Bayesian networks such as Hidden Markov Models [1–3]. They require precisely defined models that be defined before learning, but they are particularly good at capturing temporal patterns. A 3-D glove that records hand motion was utilized by Starner and Pentland in conjunction with a Hidden Markov Model (HMM) [1]. Because the glove can gather 3-D information from the hand regardless of spatial orientation, they were able to achieve an exceptional accuracy of 99.2% on the test set. Their HMM analyses time-series data to track hand motions and classify objects depending on where the hand has been in recent frames.

To translate ASL, certain neural networks have been employed [4–7]. The ability of neural networks to learn the most crucial classification features is perhaps their greatest benefit. But, to train them, they need a lot more time and information. Most have been relatively shallow up to this point. With the use of a 3-layer Neural Network and sophisticated feature extraction, Mekala et al. classified videos of ASL letters into text [4]. Hand position and movement were the two kinds of features they extracted. They note the presence of

six "points of interest" in the hand—each fingertip and the palm centre—before classifying the hand according to ASL. In order to determine which region of the frame the hand is in; Mekala et al. additionally perform Fourier Transforms on the images. Although they assert that this system can accurately classify 100% of the photos, they don't specify if this feat was accomplished on the training, validation, or test sets.

The most relevant work to date is L. Pigou et al's application of CNN to classify 20 Italian gestures from the ChaLearn 2014 Looking at People gesture spotting competition [7]. The CLAP14 dataset comprises 20 Italian sign gestures, as discussed in a paper. The author used a Convolutional Neural Network (CNN) model with six layers after preprocessing the photos. Note, the model only employs 2D kernels and is not a 3D CNN. Rectified Linear Unit (ReLU) was also used by the author as the activation function. They obtain a cross validation accuracy of 91.7% by using a Microsoft Kinect on full-body photographs of individuals making the gestures. Similar to the 3-D glove mentioned before, the Kinect enables the acquisition of depth data, which is extremely helpful for categorizing ASL signs.

In summary, the literature review shows that there have been various methods and algorithms used to tackle sign recognition in videos and images, including linear classifiers, neural networks, and Bayesian networks. High accuracy rates have been attained in several investigations using techniques like Faster R-CNN, 3D CNN, and Hidden Markov Models. The use of depth features, such as those captured by the Microsoft Kinect, also aids significantly in classifying ASL signs.

III.PROPOSED SYSTEM

We have proposed an end-to-end system. The user at the sender's side opens the application and turns on the camera. The system translates the sign language to English in real-time. On the receiver's front, English language is translated to Sign language, again in real-time. The major objective of the proposed system is to help deaf and mute people communicate effectively with people who do not comprehend sign language. The system comprises of mainly:

1. Sign language to Text translation in real-time
2. Audio/Text to Sign Language translation in real-time

In order to overcome this issue and enable dynamic communication, we have suggested an ASL recognition system that makes use of cutting-edge technology like Convolutional Neural Networks (CNN) to convert video of a user's ASL signs into text in real-time as well as generate Sign Language. These real-time activities make up our proposed solution:

1. Getting a video of the signer (input)
2. Classifying each frame in the video to a letter
3. Using categorization scores to reconstruct and present the most likely term (output)
4. Getting user-provided text or audio (input)
5. Generating Signs out of Text (output)

This subject poses a substantial difficulty for computer vision because of a variety of factors, such as:

1. Environment-related issues (e.g., camera position, background, lighting sensitivity)
2. Obstruction (e.g., an entire hand or some or all fingers can be out of the field of view)
3. Recognizing border signs

Although neural networks have been used to recognize ASL letters in the past, routinely achieving accuracy levels above 90%, many of these techniques require a 3-D capture element employing motion-tracking gloves or a Microsoft Kinect. However, only a small number of these solutions can generate sign language and offer real-time classifications, which makes the rest less scalable and practical.

Our solution is an ASL recognition method that uses a pipeline and a web application to receive video of someone signing a word as input. The system gathers individual video frames from letters A through Y and applies a CNN to create letter probabilities for each. Based on the character index that each frame is considered to belong to, the frames are grouped using a variety of heuristics. The user is then presented with a plausible term using a language model. We have proposed the creation of sign language in addition to ASL recognition. By entering text or audio and having the system generate signs from that, those who are hard of hearing or deaf would be able to effectively communicate with people who do not know sign language.

Data collection is the initial step in the suggested system. Researchers have previously utilized sensors or cameras to record hand motions. As shown in Figure 1, in the suggested method, hand gestures are captured using a webcam. The HSV colour extraction technique is used in a number of pre-processing steps to remove the background from the webcam photos. Segmentation is used to find the skin tone area once the backdrop has been removed. An elliptical kernel is used to perform a series of dilation and erosion operations after adding a mask to the pictures using morphological techniques. With OpenCV, the images are altered to the same size, eliminating any distinction between images of various gestures. A Convolutional Neural Network (CNN) is then utilized for training and categorization once the binary pixels from each frame are retrieved. After the model has been evaluated, the system then predicts the alphabet.

The real-time translation of audio/text into sign language is the second stage of the suggested system. The user has the choice to input data via text or audio on the receiver's end. The output of the natural language model's processing of the input is a list of keywords. If there are any phrases or word combinations in the keywords for which there is a sign language video in the database, those movies will be displayed; otherwise, the keywords will be tokenized further into words or letters.

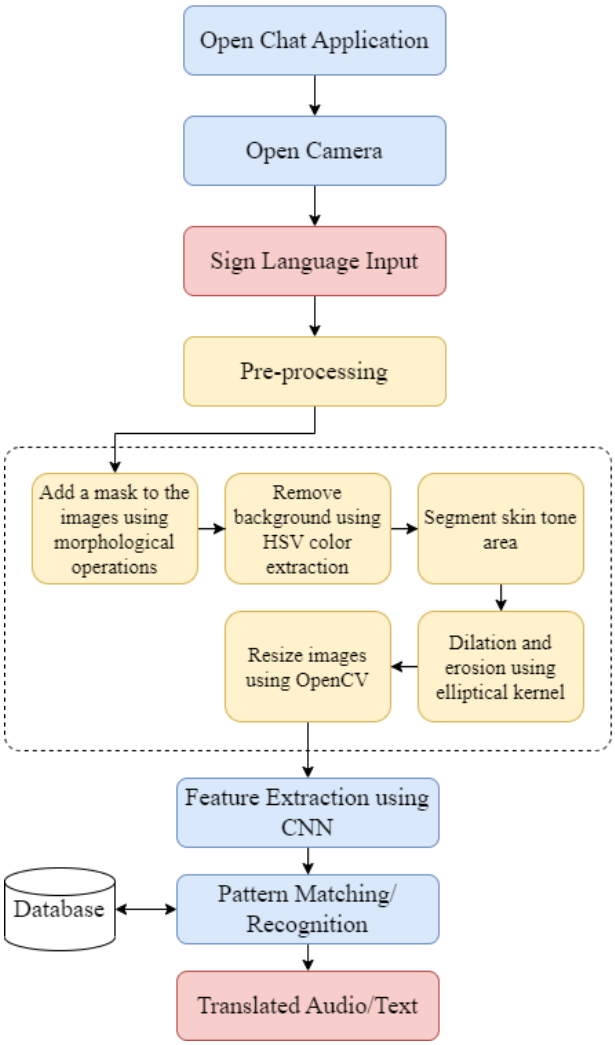

Figure 1. Flow of the proposed system: Sign Language Recognition

As shown in Figure 2, the suggested system uses the CNN as well as additional machine learning methods to enhance performance, including Hidden Markov Models (HMMs) and Support Vector Machines (SVMs). While SVMs are used to categorize sign language motions, HMMs are used to model the temporal dynamics of sign language gestures. Python is used to implement the suggested system, together with packages like OpenCV, Keras, Tensorflow, and Scikit-learn. The system is tested using a dataset of different American Sign Language movements, and the findings

show great accuracy in hand gesture recognition and translation into text, as well as the other way around.

Overall, the suggested system offers a reliable and effective method of enabling real-time communication between deaf and mute people and those who do not know sign language. The translation of sign language motions into text and audio and vice versa is accurate and dependable because to the system's usage of machine learning algorithms and computer vision techniques.

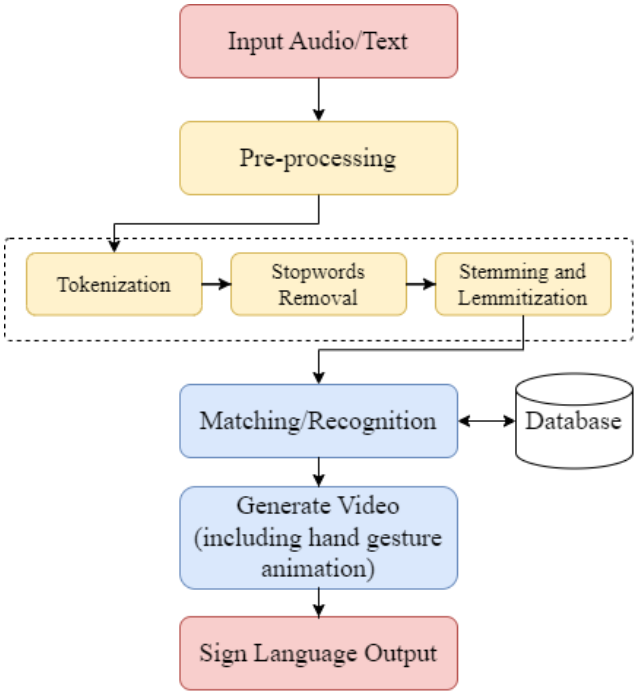

Figure 2. Flow of the proposed system: Sign Language Generation

A. Dataset

Data collecting is an important component in this study since it has a significant impact on how well the suggested system will function. For this, two datasets were used: one for sign language generation and the other for sign language detection.

7,172 photos and 785 columns make up the test dataset, whereas 27,455 images and 785 columns make up the training dataset. A 70:30 ratio between the training and test datasets indicates that 70% of the data is utilized for training and the other 30% for testing. The label of the image, which relates to the associated sign language motion or word, is found in the dataset's first column. The hand gesture is represented by a flattened 28x28 picture in the remaining 784 columns. For simple storage and processing, the image is flattened into a 784-pixel, one-dimensional array. The frequency distribution of the letters is represented in Figure 4. Note that the letters J and Z are not present in the dataset. It can be seen that there is a uniform distribution of data.

Figure 3. American Sign Language Hand Gestures

The dataset for the creation of sign language includes animations of various signs that are labelled with their English translations

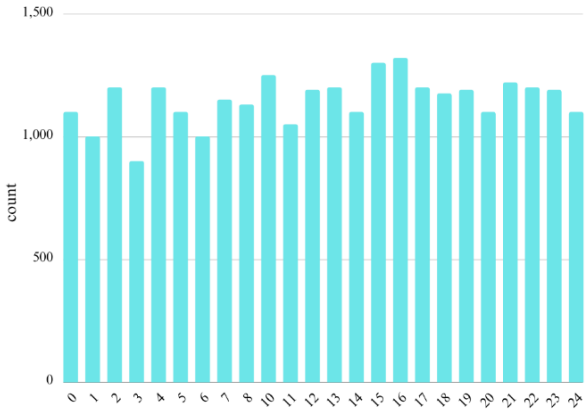

Figure 4. Count of Different Letters from the training dataset

The suggested system can accurately learn and detect different sign language motions since it has a large amount of training and testing data. The suggested technique is more useful for real-time applications since using flattened images enables quicker and more effective processing.

B. Data Processing

1) Background Elimination: It is difficult to distinguish between the hand motion based solely on skin tone because the photographs were taken in the RGB colour space. The HSV colour space is used for this image processing. HSV is a helpful tool for enhancing image stability since it can differentiate brightness from chromaticity. The Hue element can be used to erase backgrounds because it is unaffected by different types of lighting, shadows, and shading. After a track bar with H and S values of 0 to 179, 0-255, and 0 to 255 recognize the hand gesture, the background is rendered dark. Processes of elliptical kernel dilation and erosion are applied to the hand gesture's surface. Applying the two masks' results in the creation of the first image.

2) Segmentation: An essential step in image processing is the selection and extraction of key visual elements. Due to the significant quantity of data that photographs contain, when they are recorded and kept as a dataset, they frequently need a lot of storage space. By automatically removing significant characteristics, feature extraction minimizes the amount of data while preserving the accuracy of the classifier. The binary pixels of the photos were determined to be crucial features in this situation, and by scaling the images to 64 pixels, enough data were collected to correctly identify American Sign Language gestures, yielding a total of 4096 features (64x64 pixels).

A CNN model, which is a multi-layered feed-forward neural network frequently used in image recognition, is used to extract characteristics from frames and predict hand motions. The CNN architecture comprises a number of convolution layers, each of which includes a fully linked layer set, a pooling layer, an activation function, and optional batch normalization. Images are resized as they move through the network owing to max pooling. The prediction of class probabilities is produced by the top layer.

3) Classification: The system we have proposed 2D CNN model. The convolution layers scan the images using the 3 by 3 filter. The dot product of the filter weights and frame pixels is calculated. This particular stage extracts important characteristics that are then transmitted from the provided image.

C. Sign Language Generation

If the system's database does not have a video for a certain word, it can increase productivity by breaking the word down into its individual letters and showing a video output of each letter sequentially, making sure that no words are skipped. Instead of displaying each word separately, the system can also recognize phrases in sentences and display sign language videos corresponding to the complete phrase from the database. The system only has this feature.

The model first determines whether any of the movies in its database match the supplied text. The supplied text is utilized exactly as the keywords for creating the related sign language movie if a match is discovered. If no match is found, several NLP technologies—which are covered in depth under the "Methodology" heading—are used to process the input text. Because it is a potent opensource NLP library used for evaluating human language data, the NLTK library is essential to this system. Using NLTK, the text is processed in a number of ways, including tokenization, stop-word removal, lemmatization, parse tree generation, parts of speech (POS) labelling, and more.

1) Tokenization: It is the process of splitting text into words or tokens in natural language processing. It includes word tokenization and sentence tokenization.

2) Removal of stop words: Removing stop words can enhance the effectiveness of NLP tasks like topic modelling, sentiment analysis, and text categorization by eliminating common and uninformative words.

3) Parsing: This step involves analysing the syntax of a sentence to ensure it follows correct grammar and aids in modifying the text based on the language's grammar structure.

4) Lemmatization: This reduces inflected word forms to their base or dictionary form, which is crucial in Indian Sign Language for translation.

5) Parts of Speech Tagging: POS Tagging labels words with their grammatical functions, such as nouns, verbs, adjectives, adverbs, and prepositions.

IV. EXPERIMENTAL RESULTS

To test the proposed real-time audio-to-sign language and sign language-to-audio system, an experimental setup was designed. The experiment involved collecting data from deaf and mute individuals who were native sign language users. Participants were asked to use the system to communicate with a non-sign language speaker.

A. Evaluation Metric

The system's performance was assessed in terms of accuracy, speed, and usability using the participant data. The accuracy of the system was assessed by contrasting its output with the actual sign language utterances made by the participants. It was also noted how long the system took to produce the audio output and animations in sign language.

Our testing approach involved evaluating the performance of the sign language translation and generation system across a variety of different testing scenarios. For each letter in American Sign Language (ASL), we tested the system's accuracy, speed, and usability in multiple environments and conditions. This included testing the system's performance in well-lit and poorly lit environments, against different backgrounds and skin tones, and with different sign language speakers.

To measure the accuracy of the system, we compared the system's output with the actual sign language performed by the participants. This allowed us to determine the percentage of cases in which the system accurately translated or generated the sign language for each letter. This was done to ensure that the system's performance was robust and reliable across a range of real-world scenarios.

B. Results for Sign Language Recognition

A real-time CNN algorithm was used to recognize ASL fingerspelling. Deaf signs are converted into text statements in this study. By utilizing a deep learning technique, this system produced positive outcomes. The entire set of experiment-related findings are covered in this section. We conducted extensive testing. The accuracy for the training set was 98.53%, and the accuracy for the validation set was 98.84%.

Consider testing the accuracy of the letter "A" in American Sign Language (ASL) using 10 different sign language speakers, and the application accurately translated or generated the sign language for the letter in 8 out of 10 cases, then the accuracy percentage for that letter would be 80%. The maximum accuracy of 100% was attained. However, as indicated in Table 1, the letter M had the lowest accuracy (60%), while the letter N had the second-lowest accuracy (50%). The reason is that the location of the thumb is the only way to tell the letters M and N apart because they have the

same view. In addition, the system was able to correctly identify the letters J and Z despite the fact that they include motion by analysing each movement from three separate angles.

TABLE I

EACH LABEL'S PRECISION USING THE 26-CLASS MODEL

ASL Letter	Accuracy
A	100.00%
B	100.00%
C	100.00%
D	100.00%
E	100.00%
F	100.00%
G	83.33%
H	100.00%
I	100.00%
J	66.66%
K	100.00%
L	100.00%
M	83.33%
N	100.00%
O	100.00%
P	83.33%
Q	66.66%
R	100.00%
S	100.00%
T	100.00%
U	100.00%
V	83.33%
W	83.33%
X	83.33%
Y	83.33%
Z	66.66%

C. Results for Sign Language Generation

1) Case 1:
Input: How are you?
Videos Shown: {How, you?}
The user has typed "How are you" in this instance. "How you" is the sentence/keyword that results from applying NLP and ASL grammatical rules. As "are" is a stop word in this sentence, it gets eliminated. The statement is divided between the terms "how" and "you" since there isn't a video titled "How are you" in the database.

2) Case 2:
Input: Good Evening.
Videos Shown: {Good Evening}
The user entered "Good Evening" in Case 2's input field. Although the database has the video for the entire input, the user is only shown the video, which shows the phrase "Good Evening" in sign language.

3) Case 3:
Input: My name is Raj.
Videos Shown: {name, r, a, j}
My name is Raj, as entered by the user. "name Raj" is the phrase or keywords that results from applying NLP and ASL grammatical rules. Since "is" is a stop word, it is eliminated in this. "name" is displayed in a single movie since the database has a sign language video for the complete word. As the database does not have a sign language video for "Raj," it is divided by letters, and videos for each letter are displayed.

V. CONCLUSION

The Sign Language Recognition System described in this research study offers a significant step towards bridging the communication gap between the community of people with hearing and speech impairments and the wider public. The system can quickly convert dynamic gestures into either Indian Sign Language or English when they are seen in continuous visual sequences. The system is user-friendly and relies on natural language processing and grammatical conventions of American Sign Language to successfully translate between languages. The community of deaf and hard-of-hearing persons should

anticipate that the system will significantly affect their way of life. The system's integration in public locations like hospitals, post offices, railway stations, buses, and video conferencing apps will be advantageous to the community. Because of it, they will be able to interact with their surroundings more successfully and effectively. The meticulous dataset selection, the use of machine learning techniques like Convolutional Neural Networks, and the incorporation of natural language processing and grammatical norms from American Sign Language are all credited with the efficacy of the suggested system. The system has been shown to be accurate and effective in experimental results, making it a solid choice for real-world uses.

The Sign Language Recognition System, an important development in the field of assistive technology, is predicted to improve the quality of life for deaf and hard-of-hearing people. It shows how technology has the ability to increase inclusivity, accessibility, and promote equal opportunity for all people, regardless of their skills.

VI. REFERENCES

[1] Starner, T., Pentland, A.: Real-time american sign language recognition from video using hidden Markov models. Proceedings of International Symposium on Computer Vision - ISCV, 265–270 (1995)

[2] Jebali, M., Dalle, P., Jemni, M.: Extension of hidden markov model for recognizing large vocabulary of sign language. International Journal of Artificial Intelligence Applications 4 (2013) https://doi.org/10.5121/ijaia.2013.4203

[3] Suk, H.-I., Sin, B.-K., Lee, S.-W.: Hand gesture recognition based on dynamic bayesian network framework. Pattern Recognition 43, 3059–3072 (2010) https: doi.org/10.1016/j.patcog.2010.03.016

[4] Mekala, P., Gao, Y., Fan, J., Davari, A.: Real-time sign language recognition based on neural network architecture. 2011 IEEE 43rd Southeastern Symposium on System Theory, 195–199 (2011)

[5] Admasu, Y.F., Raimond, K.: Ethiopian sign language recognition using artificial neural network. 2010 10th International Conference on Intelligent Systems Design and Applications, 995–1000 (2010)

[6] Atwood, J., Farrell, J.: American sign language recognition system. (2012)

[7] Pigou, L., Dieleman, S., Kindermans, P.-J., Schrauwen, B.: Sign language recognition using convolutional neural networks, vol. 8925, pp. 572–578 (2015). https://doi.org/10.1007/978-3-319-16178-5 40

[8] He, S.: Research of a sign language translation system based on deep learning. 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), 392–396 (2019)

[9] Geetha, M.K., ManjushaU, C.: A vision based recognition of indian sign language alphabets and numerals using b-spline approximation. (2012)

[10] Herath, H.C.M., W.A.L.V.Kumari, Senevirathne, W.A.P.B., Dissanayake, M.: Image based sign language recognition system for sinhala sign language. (2013)

[11] Huang, J., Zhou, W.-g., Li, H., Li, W.: Sign language recognition using 3d convolutional neural networks. 2015 IEEE International Conference on Multimedia and Expo (ICME), 1–6 (2015)

Cite this article as :

Pusti Sheth, Ronik Dedhia, Akshit Chheda, Dr. Vinaya Sawant, "American Sign Language Recognition and Generation : A CNN-based Approach ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 4, pp.241-250, July-August-2023. Available at doi : https://doi.org/10.32628/CSEIT23902103
Journal URL : https://ijsrcseit.com/CSEIT23902103